

HIGHER ORDER MODELS AND COMPARATIVE EVOLUTIONARY ANALYSES
FOR DETECTING FUNCTIONAL DIVERGENCE AMONG GENOMIC SEQUENCES

By

ERIC A. GAUCHER

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2001

THE
FUNDAMENTALS OF
THEORY OF
THEORY OF
THEORY OF

Copyright 2000

by

Eric A. Gaucher

ACKNOWLEDGMENTS

I am very grateful to the guidance Dr. Steven Benner has provided me throughout my graduate career. Propitiously, his ability to foster my development was matched by his willingness to let me self-discover. I would like to thank M. Daniel Caraco, Ujjwal Das, Sridhar Govindarajan, David Liberles and David Schreiber for their assistance with computational analyses, and Jennifer Piascik in the Zoology Department for her assistance with drafting figures. A special thanks to the Benner group for their assistance and friendship over the years, and to Romaine for providing extraordinary management of our group.

I would also like to thank my committee members, Drs. Mike Miyamoto, Steve Sugrue and Chris West. A special thank you goes to Dr. Miyamoto for advising me as if I were one of his own students. Without his guidance, my graduate career would have been only half as rewarding as that which it actually became.

Lastly, I thank my family and Robin for providing the support that I required to achieve my ambition.

TABLE OF CONTENTS

	page
ACKNOWLEDGMENTS	iii
ABSTRACT	vi
 CHAPTERS	
1 DETERMINING THE PATHS OF DIVERGENT EVOLUTION THROUGH RECONSTRUCTED ANCESTRAL PROTEINS.....	1
Reconstructing the Divergent Evolution of Elongation Factors	1
Background on the Early Evolution of Life on Earth	1
Specific Aims, Methods and Results	8
Reconstructing the Divergent Evolution of RNases	29
Background	29
Maximum Likelihood Models for Reconstructing Ancestral RNases	32
An Alternative View of Reconstructing Ancestral Sequence "Space"	37
2 DETECTING FUNCTIONAL DIVERGENCE IN BIOLOGICAL SEQUENCES: HISTOGRAM APPROACH USING ORIGINAL RATE DIFFERENCES	40
Background	40
Methods	42
Covarian Analyses, Structural Biology, and Hypothesis Generation	44
Covarian Approaches and Functional Genomics	57
3 DETECTING FUNCTIONAL DIVERGENCE IN BIOLOGICAL SEQUENCES: HISTOGRAM APPROACH USING LOG-TRANSFORMED RATE DIFFERENCES	60
Background	60
Log-Transformation Statistics	62
4 CURRENT STATUS AND FUTURE PROSPECTS USING THE COVARION MODEL: BAYESIAN INFERENCE	72
Background	72
Evolutionary Tools for Functional Genomics	73
Covarian Approaches: Methods of Overall Sequence Comparisons	75
Covarian Approaches: Non-Bayesian-based Methods for Identifying Sites	77

Covariation Approaches: Bayesian-based Methods for Identifying Sites	80
Predictive Power of Covariation Analyses	96
Covariation Approaches, Evolutionary Tools, and Functional Genomics	99
5 EVOLUTION, LANGUAGE AND ANALOGY IN FUNCTIONAL GENOMICS ...	102
Background	102
"Functional Equivalency"	103
Orthologs as Functional Analogs?	105
A Behavioral/Functional Continuum	108
A Way Forward	110
APPENDIX	114
LIST OF REFERENCES	122
BIOGRAPHICAL SKETCH	136

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

HIGHER ORDER MODELS AND COMPARATIVE EVOLUTIONARY ANALYSES
FOR DETECTING FUNCTIONAL DIVERGENCE AMONG GENOMIC
SEQUENCES

By

Eric A. Gaucher

May 2001

Chairman: Dr. Steven A. Benner
Major Department: Molecular Cell Biology

The divergent evolution of protein sequences from genomic databases can be analyzed using different mathematical models. The most common treat all sites in a protein sequence as equally variable. More sophisticated models acknowledge the fact that purifying selection generally tolerates variable amounts of amino acid replacement at different positions in a protein sequence. In their "stationary" versions, such models assume that the replacement rate at individual positions remains constant throughout evolutionary history. "Non-stationary" covarion versions, however, allow the replacement rate at a position to vary in different branches of the evolutionary tree. Recently, statistical methods have been developed that highlight this type of variation in replacement rates. Here, we show how positions that have variable rates of divergence in different regions of a tree ("covarion behavior"), coupled with analyses of experimental three-dimensional structures, can provide experimentally testable hypotheses that relate

individual amino acid residues to specific functional differences in those branches. We illustrate this in the elongation factor family of proteins using various statistical inferences. The recent crystal structure of eukaryotic elongation factor, bound to its nucleotide exchange factor, demonstrates the predictive powers associated with incorporating the covarion model into comparative evolutionary analyses.

In addition, based on previous work in this laboratory, we show that incorporating higher order models of sequence evolution leads to predictions of ancestral reconstruction character states that are different than those predicted by the less-sophisticated parsimony method. We also lay the foundation for determining whether the ancestral organism to all extant bacteria lived in a cold- or hot-temperature environment. In conclusion, we advocate the use of higher order models in comparative evolutionary analyses especially as the community attempts to predict protein function from the large amount of genomic data currently being produced by sequencing projects.

CHAPTER 1 DETERMINING THE PATHS OF DIVERGENT EVOLUTION THROUGH RECONSTRUCTED ANCESTRAL PROTEINS

Reconstructing the Divergent Evolution of Elongation Factors

Background on the Early Evolution of Life on Earth

From 4.5 billion to 3.8 billion years ago, Earth endured a heavy bombardment from meteors. These impacts often had devastating results. It is estimated that some of the impactors were at least 500 kilometers in diameter. Impacts from such large objects would create a superheated atmosphere of vaporized rock, which would in turn have vaporized the oceans and sterilized the surface of the planet (Sleep *et al.*, 1989). Since the first signs of life appear not long after the bombardments stopped (Schopf, 1993), researchers have used the information to form two hypotheses. First, the only organisms to survive such conditions must have lived deep in the Earth's crust where temperatures were high (Sleep *et al.*, 1989). This implies that these very early organisms were thermophiles. Alternatively, life may have repeatedly been invented and extinguished by repeated sterilizing impacts (Maher and Stevenson, 1988). This implies that these very early organisms need not have been thermophiles, but could have also been mesophiles.

Meteorite blasts were just one problem facing early life. If, as proposed by Kasting (1997), energy coming to Earth from the Sun was 30% less during our planet's early years than today, the Earth's surface should have been largely frozen, at least until the Sun brightened about 2 billion years ago. This has led researchers to two more ideas. First, early life may have lived in water just beneath the ice surface. Alternatively, early life may not have survived the freeze, except by retreating to thermal vents deep in the ocean (Gaidos *et al.*, 1999).

Until 25 years ago, microbiology could not even begin to help analyze this problem, simply because the true relationship between bacteria was not correctly understood. Starting in the mid-1970's, Carl Woese (Woese and Fox, 1977) began to piece together what seemed, at first, to be a clear and coherent picture of bacterial evolution. This was done by examining rRNA sequences using phylogenetic methods.

rRNA sequences remain one of the most useful and most used of the molecular chronometers. They are present in all organisms in homologous forms. Different positions in their sequences change at different rates, allowing a researcher to not only use rRNA to determine close relationships, but also to determine distant relationships as well. Most importantly, rRNAs can be sequenced directly and rapidly by means of reverse transcriptase (Lane *et al.*, 1985).

From 16S rRNA sequences, a "Universal Tree" was built (Figure 1-1). In this tree, life on Earth was divided into three distinct domains. Prokaryotes, formerly thought to be one "kingdom", were divided by the 16S rRNA sequences into two separate lineages, eubacteria (now called bacteria) and archaebacteria (now called archaea). The most ancient point in the tree was termed the "last common ancestor," or LCA. Figure 1-2 shows the bacterial topology based on rRNAs from cultured and non-cultured strains (Pace, 1997).

Finding the oldest point (the "root") on the tree is not trivial, as the sequences themselves do not necessarily contain information about geological time. If, however, the LCA contained a pair of paralogous genes that were created by duplication before the divergence of the three domains, and if these descendents of both paralogs survive in modern representatives of each domain, then these sequences may be used to root the universal tree. Several proteins have been used to root the tree. Gogarten *et al.* (1989) used duplicated domains in H⁺ ATPases, Brown and Doolittle (1995) used aminoacyl-tRNA synthetase duplications, Baldauf *et al.* (1996) used elongation factors Tu and G, while Gribaldo and Cammarano (1998) used duplications in the signal recognition

particles to root the tree. All these studies concluded that the root of the Universal Tree lies on the branch separating bacteria from the archaea/eukaryotic bifurcation.

The Universal Tree was used to draw inferences about features of the environment of ancestral organisms, in particular, about the temperature environment of the LCA at the root of the tree. These were, for the most part, based on a combination of ideas derived from parsimony, and the notion that ancestors are more like the descendants to which they

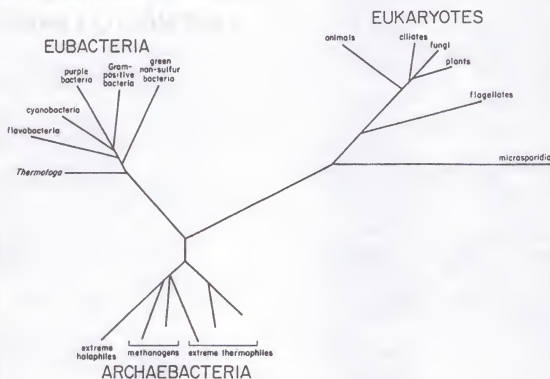


Figure 1-1. Universal Tree of Life based on 16S rRNA sequence data. Adapted from Woese (1987).



Figure 1-2. Bacterial phylogeny. Adapted from Pace (1997).

are connected by the shortest branch. For example *Aquifex pyrophilus* and *Thermotoga maritima* are both thermophiles. They both occupy branches that diverge near the base of the bacterial lineage (Burggraf *et al.*, 1992). Therefore, it was concluded that the organisms at the base of the tree were thermophiles. Similarly, the most deeply branching archaeal sequences seemed to be thermophiles (Woese, 1987). Therefore, a compelling case could be made that the LCA was a heat-lover.

Again, the idea that the LCA was thermophilic comes from the rRNA tree built by Woese and others. However, what if the rRNA tree was incorrect or not fully resolved? This is what some researchers are concluding. Delong *et al.* (1994) discovered that there are many mesophilic archaea that branch early within the archaea domain. Although these branch lengths are longer than their thermophilic counterparts, it can be argued that thermophilic organisms have a high GC content due to structural constraints, which would slow down their evolutionary rate. This decreased evolutionary rate can result in short branches. Further, Galtier *et al.* (1999) used rRNA to analyze the GC content of the

last common ancestor. Based on a maximum likelihood model that allows for variable substitution rates among lineages and assumes that the GC mutation rate has not reached equilibrium, these researchers showed that the LCA did not contain a high GC content, as appears to be necessary for thermophilicity.

Over the past few years, a number of articles have been published that call into question past interpretations. These articles question the assumption that thermophiles connected by branches near the base of the tree necessarily lead to the conclusion that the species at the base was a thermophile, discuss the fact that analyzing various genes result in different topologies for the universal tree, and the effect of lateral gene transfer. Also being questioned are the methods used to create phylogenies, the accuracy of the placement of branches deep in the tree, and a lack of the use of indel (insertion/deletion) information to test phylogenetic relationships. The next few paragraphs will further detail these ideas.

Recently, two research groups set out to study the diversity of thermophiles within different environments. Sekiguchi *et al.* (1998) used prokaryote-specific rRNA PCR primers to amplify sequences from methanogenic granular sludge. This culture-independent approach, so termed because the organism is not required to be grown in a laboratory culture, revealed interesting results. Of the 110 distinct thermophilic clones, only 22% were archaea while the remaining 78% were bacteria. Phylogenetic analysis placed some of these sequences branching as high up the bacterial lineage as the *T. thermodesulfovibrio* and green non-sulfur divisions. Hugenholtz *et al.* (1998a) performed a similar test but used bacteria-specific rRNA primers in the Obsidian Pool at Yellowstone National Park (a hot spring). Phylogenetic analysis determined that these thermophilic species branched as high up the tree as proteobacteria. Also, Hugenholtz *et al.* (1998b) concluded that several of these sequences constitute new divisions within the bacterial tree. The conclusion of the LCA being thermophilic, based on previous phylogenetic analysis solely placing thermophiles at the base of the bacterial lineage, may

no longer be strongly supported (albeit some of these results can still support ancestral thermophily).

Hasegawa and Hashimoto (1993) concluded that phylogenetic analyses based on rRNA genes could be unreliable due to extreme AT or GC nucleotide bias in the rRNA genes of some taxa. In light of the criticism for using rRNA as a phylogenetic marker, many researchers are turning to protein-coding genes for analysis of the universal tree. Examples of this are the H. Klenk and W. Zillig research groups. These researchers (Klenk and Zillig, 1994; Klenk *et al.*, 1999) used RNA polymerase sequence data to study the phylogenetic relationships of bacterial divisions. In these studies, it was found that the aquifex division grouped with the proteobacteria at the top of the tree, whereas Burggraf *et al.* (1992) showed that the branch leading to *A. pyrophilus* was placed at the base of the bacterial lineage. Also, the Klenk group concluded that mesophilic mycoplasmas were the first species to diverge from the base of the bacterial tree, thus suggesting thermophiles are not at the base of the bacterial lineage.

It is becoming increasingly apparent that many genes within eukaryotes and prokaryotes have been acquired by horizontal transfer (Jain *et al.*, 1999). Specifically, extensive horizontal transfer has taken place for operational genes (those involved in housekeeping), whereas horizontal transfer rarely takes place for informational genes (those involved in translation, transcription and other related processes) (Rivera *et al.*, 1998). Benner *et al.* (1989) suggested that the machinery for transcription, translation, and replication were present in the protogenome (the most recent common ancestor of modern life forms). Since translation is very complex and its key components tend to be universally conserved among the three domains of life, Woese (1998) argued that this complex was probably the first to be refined. Although RNA polymerases within the three domains of life share common components, there are many components that are not universal. It appears, then, transcription was refined after translation, yet it is fairly immune to lateral transfer (Jain *et al.*, 1999). Only later did genome replication become

refined. Since translation, transcription, and replication are effected by large complexes and tightly integrated, it seems intuitive that there is not much data to suggest that their components were horizontally transferred.

The problem of long-branch attraction on a phylogeny has been known for over twenty years (Felsenstein, 1978). This phenomenon results in the robust grouping of long branches on a tree, regardless of the underlying phylogeny. This occurs, for example, when an outgroup is distantly related to a set of ingroups. If any member of the ingroup evolves considerably faster than the other ingroup members, it will be placed too deeply in the tree. This ingroup is, what is termed, "attracted" (more similar) to the outgroup. Philippe and Laurent (1998) give various examples of the above situation. In particular, they show how three well-studied thermophilic divisions from the bacterial domain may be exhibiting long-branch attraction (thermus, thermotoga, and aquifex divisions). Other studies suggest that these three divisions may be more closely related to other bacterial divisions; thermus may group with cyanobacteria (Gupta and Johari, 1998), aquifex may group with proteobacteria (Klenk *et al.*, 1999), and thermotoga may group with gram+ bacteria (Gupta, 1998).

The possible monophyletic grouping between aquifex and proteobacteria was discussed in a previous section (see above). The other two possible groupings (thermus with cyano, and thermotoga with proteo) are derived from the analysis of insertion and deletion data (indels) to infer phylogenetic topologies. Here, Gupta and colleagues (1997; 1998) used insertions and deletions in Heat Shock Protein 70 (HSP70) bacterial sequences to elucidate groupings. They also used the neighbor-joining and parsimony phylogenetic methods to determine groupings. All three methods suggested that a close relationship exists between the thermus division and cyanobacteria, thus placing thermus higher up the bacterial lineage. Also, based solely on indels, they have demonstrated the possible close relationship between thermotoga and gram+ bacteria using the two proteins HSP70 and glutamate-1-semialdehyde 2,1 aminomutase. Although using indels is not a

strong statistical-based method, it is suggestive and should cause us to re-evaluate some phylogenetic relationships.

In review, the subject of thermophilicity in the last common ancestor is greatly debated and clearly unresolved. The last few years have seen a resurgence of interest in this issue due in part to new assays, methods and sequences. Hopefully novel approaches will be able to bring some clarity to this “hot” topic.

Specific Aims, Methods and Results

We considered reconstructing an ancestral protein from an organism near the base of the bacterial tree to measure its thermal stability as a way of shedding light on the issue of ancient thermophily. Any attempt to reconstruct ancestral sequences requires the use of extant sequences as the basis for reconstructions. Completely sequenced genomes were used to elucidate the best possible candidates of extant sequences. A database was generated that contained the genomes of all twelve completely sequenced organisms to date as of 09-04-98. The database was then divided into homologous (paralogous and orthologous) translated gene families based on pairwise comparisons of all genes. We identified all families of proteins that had representatives in all three kingdoms. For each family, a N-J gene tree was generated using PAM (Accepted Point Mutations per 100 amino acids) distances. A table was subsequently generated that contained a single score for all individual families. This score represented the distance (PAM value) between the two most distantly related sequences within the gene tree for a given family. For example, the lactate dehydrogenase (LDH) family had a score of 165. This means that the two most distantly related LDHs contain 165 estimated mutations per one hundred positions based on the tree for this family. However, interpretation of these scores requires some careful analysis. The size of a gene can affect PAM scores such that small genes can have more functional constraints than larger genes. This means that although a large gene could have a higher PAM score than a small gene, the small gene may have more mutations in the functional domain whereas mutations in the large gene could

congregate within loop and turn regions (nonfunctional domains). Also, a family can contain many paralogous genes that may be under different selective forces, thus resulting in a higher overall PAM score for the family due to more mutations (sequence differences) among the loci.

With this in mind, the table containing scores for each gene family was analyzed. The family with the lowest, and thus most interesting, score (61) was a hypothetical ethylene-responsive protein. However, since we are interested in testing the ancestral characteristics of an extant protein, this family is obviously of little interest because its function is unknown. The second best score (100) was from the adenylylsulfate 3-phosphotransferase family. This family contained only 6 representatives. This family also proved to be of little interest because it contained so few members. The family that represented the most potential for ancestral reconstruction was EF-Tu. This family had the thirty-fifth best score (133). It contained members from all 12 of the complete genomes. Swiss-Prot had over 80 EF-Tu sequences in its database, which would enable us to generate a phylogenetic tree that contains many branches. Also, EF-Tu has been used for many phylogenetic, structural, and biochemical studies.

Elongation Factor Tu (bacteria)/Elongation Factor 1 alpha (archaea and eukaryota) is a GTPase family member involved in cellular function. EF-Tu forms a complex with GTP that in turn favors the binding of an aminoacyl-tRNA, Figure 1-3. This ternary complex binds to mRNA-programmed ribosomes delivering aminoacyl-tRNA to the ribosomal A site (for review see Czworkowski and Moore, 1996). The correct codon-anticodon interaction alters the conformation of both the aminoacyl-tRNA and EF-Tu, by way of GTP hydrolysis. The EF-Tu/GDP complex then dissociates due to a subsequent low affinity for aminoacyl-tRNA and the ribosome. Sequences have been determined for many species and a robust tree can be constructed without large taxon gaps in it. The biochemistry of EF-Tu has been studied for over three decades resulting in a clear understanding of the functional aspects of the protein (Negrutskii and El'skaya, 1998).

Since elongation factors are involved in translation, this may presumably avoid problems associated with lateral gene transfer. EF-Tu proteins from thermophiles are thermostable whereas their mesophilic counterparts are not thermostable.

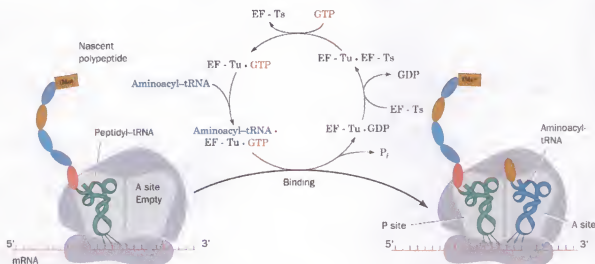


Figure 1-3. Life cycle of elongation factor. Adapted from Voet and Voet (1995)

Reconstructing a protein at the base of the universal tree requires many assumptions. To decrease some of these assumptions, we attempted to reconstruct the common ancestor of the bacterial lineage. There are two main reasons for doing this. First, reconstructing sequences at the base of any tree without an outgroup is not possible. Seeing that methods to root the universal tree are sketchy at best, using a single domain of the tree permits the other domains to act as outgroups. Second, the bacterial domain is believed to be the oldest of the three domains. The use of the bacterial lineage would enable us to have more confidence (less variability) in the reconstruction, yet go far back in time. A total of 51 complete bacterial EF-TU sequences were retrieved from Genbank and Swiss-Prot (Figure 1-4). The sequences were aligned by Darwin and ClustalW. The two programs generated highly similar alignments. Some final adjustments to the

sequences, which included minor cropping and indel rearrangements, were done by hand. The final alignment had 51 sequences with 409 positions.

The next step in the analysis was to generate a tree topology using the aligned sequences. This can be done using any one, or combination, of a number of methods. The most commonly used methods are the parsimony, distance and maximum likelihood approaches. Although the maximum likelihood method is regarded as the most statistically grounded approach (see below), it would take too much computational time to resolve a topology for 51 sequences. One way around this problem is to narrow the "tree space" that maximum likelihood must search to find the best topology. This is accomplished by giving maximum likelihood a constrained tree. A constrained tree forces some taxa to group together whereas other taxa are free to form various groupings or relationships. The parsimony, distance and "approximate" maximum likelihood methods were used to generate a constrained tree.

The principle of parsimony is quite simple. This method searches for a tree that requires the smallest number of evolutionary changes to explain the differences observed among the taxa. Parsimony relies on what are known as informative sites. A site is considered phylogenetically informative only if it favors some trees over other trees. Given a topology, one can count the minimum number of replacements (between extant and ancestral nodes) to generate the tree value. Therefore, the most parsimonious tree is the topology, or set of topologies, that contains the fewest number of substitutions (Fitch, 1971). Unfortunately computer simulations have shown that this method is less reliable when there is variation in branch lengths and when multiple substitutions have taken place at a given site (Tateno *et al.*, 1994). This computationally quick method works best to determine simple evolutionary relationships. PAUP 4.0 beta version was implemented to find such relationships among the taxa (Swofford, 1998). Bootstrap analysis (Felsenstein, 1985) using parsimony revealed many statistically relevant relationships

OTU	Species Name	Entry	Name	SWISS-PROT Primary accession
Spirochetes				
TPAL	<i>Treponema pallidum</i>	O83217		EFTU_TPRA
BBUR	<i>Borrelia burgdorferi</i>	P50062		EFTU_BORBU
TREHY	<i>Treponema hyodysenteriae</i>	P52854		EFTU_TREHY
Thermotoga				
THEMA	<i>Thermotoga maritima</i>	P13537		EFTU_THEMA
FERTS	<i>Fervidobacterium islandicum</i>	O50340		EFTU_FERTS
Deinococci/Thermus				
THETH	<i>Thermus aq. (thermophilus)</i>	P07157		EFTU_THETH
THEAQ	<i>Thermus aquaticus</i>	Q01698		EFTU_THEAQ
DEISP	<i>Deinonema sp</i>	P33168		EFTU_DEISP
Bacteroides				
BACFR	<i>Bacteroides fragilis</i>	P33165		EFTU_BACFR
CYTLY	<i>Cytophaga lytica</i>	P42474		EFTU_CYTLY
TAXOC	<i>Taxobacter ocellatus</i>	P42480		EFTU_TAXOC
FLAFE	<i>Flavobacterium ferrugineum</i>	P42476		EFTU_FLAFE
CHLVI	<i>Chlorobium vibrioforme</i>	P42473		EFTU_CHLVI
Gram Positive (Low G+C)				
BACSU	<i>Bacillus subtilis</i>	P33166		EFTU_BACSU
BACST	<i>Bacillus stearothermophilus</i>	O50306		EFTU_BACST
MYCGA	<i>Mycoplasma gallisepticum</i>	P18906		EFTU_MYCGA
MGEN	<i>Mycoplasma genitalium</i>	P13927		EFTU_MYCGE
MPNPU	<i>Mycoplasma pneumoniae</i>	P23568		EFTU_MYCPN
UREUR	<i>Ureaplasma urealyticum</i>	P50068		EFTU_UREUR
MYCHO	<i>Mycoplasma hominis</i>	P22679		EFTU_MYCHO
STROR	<i>Streptococcus oralis</i>	P33170		EFTU_STROR
HPYL	<i>Helicobacter pylori</i>	P56003		EFTU_HELPI
Cyanobacteria				
ANANI	<i>Anacystis nidulans</i>	P18668		EFTU_ANANI
SYNP7	<i>Synechococcus sp.</i>	P33171		EFTU_SYNP7
SYNP3	<i>Synechocystis sp.</i>	P74227		EFTU_SYNP3
SPIPL	<i>Spirulina platensis</i>	P13552		EFTU_SPIPL
Proteobacteria (purple)				
AGRTU	<i>Agrobacterium tumefaciens</i>	P75022		EFTU_AGRTU
STIAU	<i>Stigmatella aurantiaca</i>	P42479		EFTU_STIAU
RICPR	<i>Rickettsia prowazekii</i>	P48865		EFTU_RICPR
BURCE	<i>Burkholderia cepacia</i>	P33167		EFTU_BURCE
THICU	<i>Thiobacillus cuprinus</i>	P42481		EFTU_THICU
NEIGO	<i>Neisseria gonorrhoeae</i>	P48864		EFTU_NEIGO
SALTY	<i>Salmonella typhimurium</i>	P21694		EFTU_SALTY
ECOLI1	<i>Escherichia coli</i>	P02990		EFTU_ECOLI
ECOLI2	<i>Escherichia coli</i>	P02990		EFTU_ECOLI
HINF	<i>Haemophilus influenzae</i>	P43926		EFTU_HAEIN
SHEPU	<i>Shewanella putrefaciens</i>	P33169		EFTU_SHEPU
WOLSU	<i>Wolinella succinogenes</i>	P42482		EFTU_WOLSU
CAMJE	<i>Campylobacter jejuni</i>	O69303		EFTU_CAMJE
Gram Positive (High G+C)				
MICLU	<i>Micrococcus luteus</i>	P09953		EFTU_MICLU
BRELW	<i>Brevibacterium linens</i>	P42471		EFTU_BRELW
CORGL	<i>Corynebacterium glutamicum</i>	P42439		EFTU_CORGL
PLARO_A	<i>Planobispora rosea</i>	X98830		EFTU_PLARO
PLARO_B	<i>Planobispora rosea</i>	U67308		
STRJC	<i>Streptomyces cinnamonensis</i>	P95724		EFTU_STRJC
STRAU	<i>Streptomyces aureofaciens</i>	O33594		EFTU_STRAU
MTUB	<i>Mycobacterium tuberculosis</i>	P31501		EFTU_MYCTU
MYCLE	<i>Mycobacterium leprae</i>	P30768		EFTU_MYCLE
Aquifex				
AQUFY	<i>Aquifex pyrophilus</i>	O50293		EFTU_AQUFY
AQUAE1	<i>Aquifex aeolicus</i>	O66429		EFTU_AQUAE
AQUAE2	<i>Aquifex aeolicus</i>	O66429		EFTU_AQUAE
Eukaryota				
GLUPL	<i>Glugea plecoglossi</i>	D32139		
Archaeobacteria				
METVA	<i>Methanococcus vannielii</i>	P07810		EF1A_METVA
SULAC	<i>Sulfolobus acidocaldarius</i>	P17196		EF1A_SULAC

Figure 1-4. List of taxa names used in the Elongation factor analyses.

within the data set. Although this method generated many relationships among the taxa (data not shown), it did not establish enough relationships to generate a constrained tree suitable for a maximum likelihood analysis.

The approximate maximum likelihood method (Adachi and Hasegawa, 1996) was used to determine more relationships among the bacterial taxa. This method first generates a Neighbor-Joining (N-J) distance tree and then uses the maximum likelihood approach to search the tree space around the N-J tree (see below for explanation on both distance and maximum likelihood methods). This analysis works on the assumption that the N-J tree tends to be very similar to the maximum likelihood tree. Unfortunately we are not guaranteed to find the best tree using this method. It is however a good method to generate relationships from sequences that are more distantly related than those generated by parsimony. The REL Bootstrap analysis method using approximate maximum likelihood as implemented in Molphy was used on the EF data set. The analysis generated many evolutionary relationships (data not shown), in addition to those generated by parsimony.

For a variety of applications, evolutionary distances between sequences are needed. Distance measures are generally constructed so that the distance between two sequences scales linearly with the number of mutations that occurred during the evolutionary history separating them. Simple tools for estimating these distances assume that all sites in the sequence change at the same rate. In fact, a simple inspection of any multiple sequence alignment will show that this is not the case. Some positions are more variable than others. This fact, as it turns out, makes distances calculated using these simple tools inaccurate. Modeling this rate heterogeneity among sites can overcome these associated problems.

Rate heterogeneity is a concept that dates back to 1967 when Fitch and Margoliash were analyzing cytochrome *c* sequence data. They discovered that the number of substitutions per site did not follow a Poisson distribution. Up to this time, researchers

proposed that all sites in a sequence change at the same rate (rate homogeneity) and therefore conform to the Poisson distribution. Further analysis revealed sites tend to follow a negative binomial (Golding, 1983). This so-called gamma distribution (rate heterogeneity) more accurately models rate variation. To better understand rate variation among sites consider two sets of sequences, each containing two DNA sequences. The first set contains sites that are allowed to mutate anywhere along the sequence. The second set contains sites in which only 90% of the sequences can mutate; therefore they always have 10% identity. Now, after infinite time, the first set will contain 25% sequence identity due to random chance using four nucleotides. After infinite time and the same mutation rates as the first set, the second set will have 32.5% sequence identity ($0.25 \times 0.90 + 0.10$). Estimates of evolutionary distances can be biased, and therefore flawed, unless this rate heterogeneity is taken into account. In reality, unlike this simple example, substitution rates are estimated from sequence similarity data. It is not possible to generate a true tree topology if the number of substitutions between sequences is incorrectly calculated. Yang (1996) gives a nice review of the gamma distribution and its main parameter alpha. A large alpha represents little rate variation (some pseudogenes) and a small alpha value represents extreme rate variation (globin genes). Sullivan *et al.* (1996) showed that the estimate of alpha based on a well-corroborated tree topology is well outside the distribution of estimates derived from random trees. However, Yang *et al.* (1995a) demonstrated that it is possible to accurately estimate alpha as long as the given topology is a rough approximation of the true tree. A well-analyzed topology based on RNA sequences (Pace, 1997) (Figure 1-2) was used to estimate the EF-Tu alpha value using Yang's PAML program. An alpha value of 0.47 was obtained. Alpha was also estimated within two different conditions. First, some of the branches were shuffled around in the Pace topology. Second, a different topology than Pace's (Baldauf *et al.*, 1996) was used on the data set and contained only 13 of the 51 sequences. Both of these alternative methods generated alpha estimates between 0.48-0.50.

Distance methods can have advantages over parsimony for a number of reasons. For example, the methods can more easily account for multiple substitutions at sites and they can better resolve a topology with many varying branch lengths. These are integral to the EF analysis because analyzing very ancient divergences can encounter many lineages with substantially different branch lengths. To create a distance tree, a distance matrix based on the taxa is generated and then the taxa are “joined” together. A distance matrix is generated by pairwise sequence comparisons of the given taxa. Each sequence is compared to every other sequence and branch lengths (substitutions) are estimated for each pair. The most commonly used approach is the least-squares method (Rzhetsky and Nei, 1993). Although this method is not computationally difficult, it is laborious to explain. The reader is referred to Fitch and Margoliash (1967) for a detailed explanation (their method is the same as the least-squares method when one uses five or fewer taxa). Such methods result in the pairwise distances between sequences. The distances have been corrected for multiple hits and are therefore not the observed distances. The next step is to form a topology based on the distance matrix values. The most common method to generate a topology is the Neighbor-Joining method. However, the N-J method is just an approximate method of the Minimum Evolution method (ME). For all alternative trees it is possible to estimate the lengths of each branch from the estimated pairwise distances and then calculate the sum (S) of all branch-length estimates. The minimum evolution criterion is to choose the tree with the smallest value of S (Cavalli-Sforza and Edwards, 1967). However, since it is too computationally intense to calculate the S value for all trees, a N-J tree is initially generated and then the minimum evolution criterion is used to search the tree space around the N-J tree. Instead of creating a distance matrix using the least-squares method, we generated a matrix using a maximum likelihood method that incorporated the alpha value of 0.47 (Kumar *et al.*, 1993). This distance matrix was subsequently used in PAUP 4.0 beta to generate minimum evolution topologies for all the bacterial divisions.

The consensus topologies for all nine bacterial lineages, generated by the parsimony, approximate ML, and ME approaches, were used to create a constrained tree (Figure 1-5). Thus, the number of taxa have essentially decreased from 51 to nine. In mathematical terms, the number of possible tree topologies decreased from roughly 3×10^{76} to about 3.5×10^7 . The constrained tree was subsequently used for a full-blown maximum likelihood analysis.

Maximum likelihood is considered a higher-order statistical analysis because it incorporates an explicit probabilistic model for substitution processes (Felsenstein, 1981). Maximum likelihood calculates the transition probability from one residue to another in a time interval for each branch. This requires a substitution matrix. The most commonly used substitution matrix for protein data is the MDM published by Dayhoff (1978). The MDM was calculated from a study of the exchange probabilities derived from an analysis of the evolutionary changes seen in groups of very similar proteins. The ML method can then use these probabilities to calculate the likelihood score of each ancestral site incorporating various parameters (e.g., gamma distribution or transition/transversion ratios). The product of the likelihoods for all the sites is then computed and this result is the likelihood score. For a more thorough explanation see Felsenstein (1988), Hasegawa *et al.* (1991), and Li and Gouy (1991).

Yang's PAML program was used to perform the maximum likelihood analysis (PAML is the only ML program that can use the gamma distribution for amino acid data). The estimated likelihood score for the EF-Tu data based on the Pace topology without using alpha was -14967. The estimated likelihood score using an alpha of 0.5 was -13685. Since these two tree scores are highly significantly different ($p < 0.005$) according to the log likelihood ratio test, it is important that alpha is incorporated into the analysis. Unfortunately, PAML does not have a good tree-searching algorithm. Therefore a combination of the Molphy

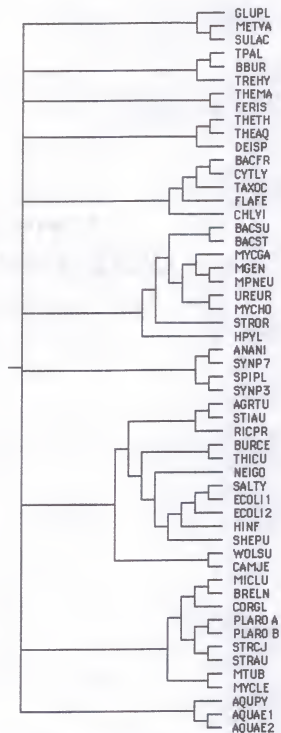


Figure 1-5. Consensus Tree generated from the analysis of individual bacterial divisions using the parsimony, distance, and approximate likelihood methods.

and PAML programs was implemented to analyze the data using the constrained lineages. First, Molphy searched for tree topologies. This analysis utilized the approximate ML method incorporating the JTT substitution model (Jones *et al.*, 1992) (similar to the Dayhoff matrix). The top 2000 trees were saved and then used in the full-blown ML method in Molphy. This generated a series of likelihood scores with bootstrap values. The top 15 trees (according to likelihood scores, bootstrap values and biological relevance) were submitted for PAML analysis with the alpha value set to 0.47.

Unfortunately, REL bootstrap scores revealed that no single topology significantly fit the data better than any other topology. However, four of the 15 topologies did make up the vast majority of the bootstrap scores (~90%). These four topologies were very similar to the Pace topology (in fact, one was the Pace topology). Thus, we may be able to accept the Pace topology as representing the true relationships among bacterial divisions.

In 1995, Yang, Kumar and Nei devised a new model-based likelihood method for reconstructing ancestral sequences. This method followed a standard statistical theory: that given the data at a site, the conditional probabilities of different reconstructions can be compared and the reconstruction having the highest conditional probability is the best estimate (Yang *et al.*, 1995b). This new method was superior to parsimony for two reasons. First, it used a probability-of-substitution matrix to calculate the chance of all various substitutions between residues. Secondly, it incorporated branch lengths into its estimates of reconstruction probabilities. This method was also unique as a likelihood approach because other likelihood models regarded ancestral character states as random variables (Felsenstein, 1981; Goldman, 1990). The idea of reconstructing ancestral sequences dates back almost thirty-five years (Pauling and Zuckerkandl, 1963).

Since the early part of this decade a number of research groups have used this idea to study the chemical and physiological properties of ancient proteins (Stackhouse *et al.*, 1990; Malcolm *et al.*, 1990; Adey *et al.*, 1994; Jermann *et al.*, 1995). Seeing that these

studies all used the parsimony method to infer their ancestral sequences, it is a new approach to use likelihood to infer the ancient sequences.

The Pace topology was used with the EF-Tu data to reconstruct ancestral sequences. Figure 1-6 shows a region of the ancestral sequence at the ancestral node of all bacteria using two archaea and one eukaryote as outgroups. Using 75% probability as the cutoff, the ancestral sequence has 36 ambiguous residues out of 392 total sites. Based on crystal structure data of the *E. coli* and *Thermus aquaticus* EF-Tu (Nissen *et al.*, 1995; Polekhina *et al.*, 1996; Kawashima *et al.*, 1996) ca. 17 of the 36 ambiguous residues lie in regions of the protein that do not have secondary structures of helices or strands (Figure 1-7). Also, the ancestral sequence contains, with high probability, the residues that have been deemed necessary for proper EF-Tu function (Harmark *et al.*, 1990; Cool and Parmeggiani, 1991; Weijland and Parmeggiani, 1993; Cetin *et al.*, 1998). Therefore, this topology gives rise to a sequence (even if we ignore ambiguities) that could be functionally active if synthesized in the laboratory.

The PAML-computed ancestral sequences can be constructed in the laboratory using the Splicing-by-Overlap-Extension PCR (SOE-PCR). Since it is not possible to know which of the sequences is the true ancestor of the bacterial lineage, all possible combinations of residues at ambiguous sites in the proteins should be generated. This requires the construction of about 2^{36} , 7×10^{10} , sequences. However, this is the case only when there are two possible residues at an ambiguous site. The ancestral sequence based on Pace's topology contains a few sites with three possible residues, thus requiring the construction of a little more than 2^{36} sequences. This can be accomplished by synthesizing primers that can generate all possible ambiguous residues in the sequence. PCR methods could subsequently mutate a given sequence (whichever extant sequence most closely resembles the ancestral sequence) to result in all possible combinations of ancestral sequences (see section on *Bacillus* below for detailed explanation).

```

Site  Freq  Data:
229  1  EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEIII: E(1.000)
230  1  DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDQ: D(1.000)
231  1  VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVSD: V(1.000)
232  1  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFMMFFIVV: F(1.000)
233  1  STSSSTSSSTTTSSSTTTSSSSSSSTSTSTSTSSSSSSSTTSTHY: S(1.000)
234  1  IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIKTS: I(1.000)
235  1  STSTTTTSSSTTTTSTTTTSTSSSTATTTTTTSPASSSSSSATTITTSII: A(0.001)
    N(0.000) K(0.000) S(0.029) T(0.969)
236  1  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGPTS: G(1.000)
237  1  RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRGGG: R(1.000)
238  1  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGIV: G(1.000)
239  1  TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGG: T(1.000)
240  1  VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVMT: V(1.000)
241  1  VAVAAVVVGA VAVVVVVVVVVVVVVSAVVA VAVVVVVVVVAAVVV: I(0.000)
    V(1.000)
242  1  TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTYP: T(1.000)
243  1  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGT: G(1.000)
244  1  RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRGG: R(1.000)
245  1  VIVIVVVIIIVIIIVAVVVIIIVVVI IIIIIIVIVVIVVVVIIIRR: I(0.792)
    M(0.000) V(0.208)
246  1  EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEVV: E(1.000)
247  1  RRRTRRRRRRTTTRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRCSE: R(1.000)
248  1  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTS: G(1.000)
249  1  ISVTVQIVITVVVRKITQVIVVIIIIKRVVTKRKIVVVIIEVKS: A(0.005)
    R(0.001) N(0.000) C(0.000) Q(0.000) E(0.000) G(0.000) I(0.205)
    L(0.006) K(0.000) H(0.007) F(0.000) P(0.000) S(0.001) T(0.003)
    V(0.771)
250  1  VVLIIVLVILAVIIVVLLIVVIVVILLVIVVILVLLIIIIIVLLIV: A(0.000)
    I(0.420) L(0.104) M(0.008) F(0.000) T(0.000) V(0.466)
251  1  KKRHKLKKNNKKKKQNHKKKKRKRKKKKRKRKKKKKKKKKSIIL: R(0.589)
    Q(0.001) H(0.000) K(0.409)
252  1  VVPVVVPVVVVPVVPVIVLVVVVVVVVVVVPVVPVVPVVVVVLEKK: A(0.005)
    Q(0.000) I(0.000) L(0.001) P(0.985) S(0.001) T(0.001) V(0.006)
253  1  GGGGGGNGGNGGQGGQGNNGNNGGGGNGGGGGGGGGGGGNGGPPV: G(1.000)
254  1  EEDDDDDDDDDDDDDSEEDDEEDDEEDDDDDDDQEDDQEEEGGG: N(0.001)
    D(0.996) E(0.003)
255  1  ETEPEETEDAEETETEEEEQEETETTEPEEEEEEEEEEEEEETEMD: D(0.002)
    Q(0.000) E(0.998)
256  1  VIVIVVIIIVVAVIIVVIVVIVVVVVVVVVVVVVVVVVVVVIVVVV: A(0.000)
    I(0.007) V(0.993)
257  1  EEEEEEEEEDEEEEEEEEEDEEEEEDEEEEEEEEEEEEEEEEEEE: E(1.000)
258  1  IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIK: I(1.000)

```

Figure 1-6. Example of ancestral sequence output generated by PAML for the ancestral node of the complete bacterial lineage. “Site” refers to the amino acid position in the protein sequence. “Frequency” refers to the number of times that the given data string for that site appears in the protein sequence. “Data” refers to the residue of all 51 sequences at the given site. The probabilities for the ancestral residues are given after the data (residues given in single letter amino acid code).

	1	21	51
<i>E. coli</i>	MSKEKFERTK	PHVNVGTIGH	VDHGKTTTLA AITTVLAKTY G-GAARAFDQ IDNAPEEKAR
<i>Thermus</i>	MAKGEFIRTK	PHVNVGTIGH	VDHGKTTTLA ALTYVAAAN PNVEVKDYGD IDKAPEERAR
	*: * *	*****	*: * * *: : *: : *: : *
<i>E. coli</i>	EEEEEEEE	?	? ? ? ?
<i>Thermus</i>	E EEEEEEEE	HHHHH HHHHHHHHH	HHHHH HHHHH
	61	81	111
<i>E. coli</i>	GITINTSHVE	YDTPTRYAH	VDCPGHADYV KNMITGAAQM DGAILVVAAT DGPMPQTRH
<i>Thermus</i>	GITINTAHVE	YETAKRHYSH	VDCPGHADYI KNMITGAAQM DGAILVVSAA DGPMPQTRH
	*****:*	*:..*:*	*****: *****: *****:*
<i>E. coli</i>	EEE EEEEEEEEE	EEE HHHHH HHHHH	EEEEEEEE HHHH
<i>Thermus</i>	EEE EEEEEEEEE	EEE HHHHH HHHHH	EEEEEEEE HHHHH
	121	141	171
<i>E. coli</i>	ILLGRQVGVP	YIIVFLNKCD	MVDDDEELLE VEMEURELLS QYDFPGDDTP IVRGSALKAL
<i>Thermus</i>	ILLARQVGVP	YIVVFNNKVD	MVDDPELDDL VEMEVRDLLN QYDFPGDEVP VIRGSALLAL
	:**	*:***:*	*****:*****:*****:*
<i>E. coli</i>	?	?	??
<i>Thermus</i>	HHHHHH E EEEEEEEE	HHHHH HHHHHHHHH HH	E EEEEEHHHHH
	HHHHHH E EEEEEEEE	HHHHH HHHHHHHHH HH	EE EEEEEHHHHH
	181	201	231
<i>E. coli</i>	E-----	-GDAEWEAKI	LELAGFLDSY IPEPERAIDK PFLLPIEDVF SISGRGTVV
<i>Thermus</i>	EEMHKNPKTK	RGENEWVDKI	WELLDALDEY IPTPVRDVKD PFLMPVEDVF TITGRGTVAT
	*	*: * *	*: * * *: : *: : *: : *
<i>E. coli</i>	???? ?	?	?
<i>Thermus</i>	H HHHHH HHHHHHHH	EEEEEEEE EEEEEEEE	EEEEEEEE EEEEEEEE
	HHHHH HHHHHHHH	EEEEEEEE EEEEEEEE	EEEEEEEE EEEEEEEE
	241	261	291
<i>E. coli</i>	GRVERGIIVK	GEEVEIVGIK	-ETQKSTCTG VEMFRKLLE GRAGENVGLV LRGIKREEIE
<i>Thermus</i>	GRIERGKVKV	GDEIVEIVGLA	PETARKTVVTG VEMHRKTLQE GIAGDNVGLL LRGVSRREEV
	:*	*:**:	**::* ** *:*:* * *:*:*:
<i>E. coli</i>	??	?	?
<i>Thermus</i>	EEE EEEEE EEEEEEE	EEEEEEEE EEEEEEEEE	EEE EEEEE EE
	EEEEEEEE EEEEEEEE	EEEEEE EEEE E	EEEE EE
	301	321	351
<i>E. coli</i>	RQGVLAKEPGT	IKPHTKFESE	VYILSKDEGG RHTPFFKGYR PQFYFRITDV TGTIELPEGV
<i>Thermus</i>	RQGVLAKEPGS	ITPHTKFEAS	VYILKKEEGG RHTGFTTGYR PQFYFRITDV TGVVRLPQGV
	*****:*	*:*****:	*****:*****:*****:*
<i>E. coli</i>	EEEE	EEEE	EEEE E EEEEEEEEE EEEEE
<i>Thermus</i>	EEEE	EEEEEEEE	EEEE EEEEEEEEE EEEEE
	361	381	401
<i>E. coli</i>	EMVMPGDNIK	MVVTILHPIA	MDDGLRFAIR EGGRTVGAGV VAKVLS
<i>Thermus</i>	EMVMPGDNVT	FTVELIKPVA	LEEGLRFAIR EGGRTVGAGV VTKILE
	*****:*	*:*****:	*****:*****:*****:*
<i>E. coli</i>	EEEE EEEEE EEEEE	EEEEEE EEEEEEEEE	EEE
<i>Thermus</i>	EEEE EEEEEEEEE	EEEEEE EEEEEEEEE	EEEE

Figure 1-7. Alignment and secondary structure of *E. coli* and *T. aquaticus*. Question marks indicate ambiguities in the bacterial ancestral sequence. Asterisks indicate sequence identity, two dots indicate high sequence similarity, and one dot indicates moderate sequence similarity. "H" indicates helix and "E" indicates strand.

The PCR products can be cloned into the expression vector pET24C (Novogen). This vector provides a His-tag to facilitate purification of the expressed protein on a nickel-nitrilotriacetic acid column (Worix *et al.*, 1995). The resulting plasmid could then be transformed into the strain of *E. coli* HB101. This would result in the expression and subsequent purification of the proteins. Seeing that the ancestral sequences may be thermophilic, we wondered as to whether an exogenous thermophilic protein could be isolated from a mesophilic host. Not only is the answer yes, in addition the answer is yes for EF-Tu. Orsola Tiboni and colleagues (Tiboni *et al.*, 1989; Sanangelantoni *et al.*, 1996) have studied EF-Tu in *Thermotoga maritima*. *T. maritima* is a bacterial species with an optimal growth temperature between 80-85°C. These researchers have isolated *T. maritima* EF-Tu that was expressed in *E. coli*. They showed that the protein was active through a series of GDP/GTP assays. Another research group consisting of Mathias Sprinzl and colleagues has done the same work using *Thermus thermophilus* EF-Tu (Ahmadian *et al.*, 1991; Blank *et al.*, 1995; Nock *et al.*, 1995). *T. thermophilus* is a bacterial species with an optimal growth temperature at 75°C. In addition, the archaeal thermophilic EF-Tu from *Sulfolobus solfataricus* has been expressed and purified from *E. coli* (Masullo *et al.*, 1997). These results all indicate that the ancestral sequences can be isolated whether or not they are thermophilic.

Once the PCR products are cloned and the library of ancestral sequences is expressed, the thermostability of the proteins can be tested. This is accomplished through a series of assays; binding of GDP and GTP, and testing the intrinsic GTPase activity of EF-Tu (Fansano *et al.*, 1982; Tiboni *et al.*, 1989; Ahmadian *et al.*, 1991; Masullo *et al.*, 1994). The nitrocellulose filter method can be used to test EF-Tu's ability to bind nucleotides (Arai *et al.*, 1972; Masullo *et al.*, 1991). The reaction mixtures can contain a solution of 20 mM Tris/HCl pH 7.8, 10 mM MgCl₂, 50 mM KCl, 7 mM 2-mercaptoethanol, 50 µM [³H] GDP and 2 µM ancestral EF-Tu. The reactions can be incubated at various temperatures, filtered through a nitrocellulose membrane, washed, and the radioactivity

determined. The same procedure can be used for [^3H] GTP. These two sets of experiments would yield data that in turn are graphed out to give a curve (temperature on the x-axis and % nucleotide bound on the y-axis) that would indicate the optimal nucleotide binding temperature.

EF-Tu hydrolysis of GTP requires the presence of ribosomes and aa-tRNA (Kaziro, 1978). However, it has been demonstrated that EF-Tu can hydrolyze GTP in the presence of kirromycin or mono/divalent cations (Fasano *et al.*, 1978; Fasano *et al.*, 1982; Masullo *et al.*, 1994). Since kirromycin is unstable at high temperatures, it is best to use cations. Based on the above research, K^+ and Na^+ monovalent cations have the greatest effect on stimulating the intrinsic EF-Tu GTPase activity. For example, the assay can proceed as follows: 20 mM Tris/HCl pH 7.8, 1 mM dithiothreitol, 10 mM MgCl_2 , 3.6 M NaCl, 0.5 μM EF-Tu and 50 μM [$\gamma\text{-}^{32}\text{P}$] GTP. The reactions are allowed to incubate at various temperatures. The reactions are stopped by the addition of HClO_4 and this results in the formation of phosphododecamolybdate. This complex is then extracted with isopropyl acetate and an aliquot of the organic phase is dried on filter paper. The radioactivity is subsequently measured by a scintillation spectrometer (Fasano and Parmeggiani, 1981). Again, a graph can be generated showing the optimal temperature of GTP hydrolysis for the ancestral sequences.

The previously mentioned assays enable one to infer the intracellular temperature at which the tested ancestral EF-Tu sequence functioned. This can in turn be used to make assumptions regarding the temperature of the environment that the bacterial ancestor lived in. However, the major difficulty with these sets of experiments is not knowing which of the generated sequences is the true ancestral sequence. This can hopefully be overcome by calculating two statistical distributions. Consider the following. Assume that we know the ancestral sequence from which all bacterial divisions are derived (mesophilic and thermophilic divisions). We also know that the sequence is thermophilic. We are interested in knowing the effect mutations have on the

thermophilicity of the protein. It is possible to analyze this problem in two ways. First, we might randomly mutate the ancestral sequence and see what fraction of the mutants remain thermostable. Second, we might mutate only those positions in the ancestral sequence that are different from the extant sequences and test thermophilicity. Obviously the latter approach is more desirable because we are mutating positions within the same evolutionary space/path as the mutations that resulted in change from the ancestor to the extant species. This approach will generate a distribution of temperatures because some mutations will decrease thermophilicity whereas others may increase or cause no change. In this example a distribution was generated where the ancestral sequence was assumed to be thermophilic. The same set of experiments can be performed with the ancestral sequence being mesophilic. Again, a distribution of temperatures is generated based on mutations within a known evolutionary space/path. The two distributions between the thermophilic and mesophilic ancestral sequence examples should be different. The complexity of protein dynamics may lead one to believe that it's simpler to mutate a thermophile into a mesophile than vice versa. The mesophilic ancestor's distribution will fall more around the true temperature, whereas the thermophilic ancestor's distribution will be skewed more towards mesophilic temperatures. Thus, one could generate EF-Tu ancestral sequences at the base of the bacterial lineage, test thermophilicity, create a distribution, and compare this distribution to the aforementioned distributions to extrapolate whether the ancestral sequence was mesophilic or thermophilic. The question then becomes how will the two standard distributions be generated from which the sequences can be compared.

Based on the same principles as the above examples, it is possible to take two closely related sequences, where one is mesophilic and the other is thermophilic, reconstruct the ancestral sequence and generate a distribution. Two possible examples of this lie within the *Bacillus* and *Methanococcus* genera. Each of these two genera contains two closely related sequences with one sequence being mesophilic and the other being thermophilic.

Fortuitously, the *Bacillus* ancestor is generally believed to be mesophilic, whereas the *Methanococcus* ancestor is generally believed to be thermophilic. Therefore the two types of standard distributions can be generated.

The *Bacillus subtilis* (optimal growth temperature 35-40°C) and *Bacillus sterothromophilus* (optimal growth temperature 60-65°C) EF-Tu's have been cloned and sequenced (Ludwig *et al.*, 1990; Krasny *et al.*, 1998). Based on Pace's topology, the probabilistic ancestral EF-Tu sequence for these two species has been generated by PAML. Using 75% probability as a cutoff, the ancestral reconstructed sequence contains 16 ambiguous positions, or roughly 65,000 possible sequences. Six of the ambiguous sites are in the same positions as the ancestral sequence generated at the base of the bacterial lineage. The sixteen ambiguous sites fall into eleven distinct sites or clusters. For example, two sites are located at positions 6 and 8, one site is located at position 73, and three sites are located at positions 183, 189, and 193. Since it is not known which are the correct residues at the ambiguous positions, one needs to generate sequences with all possible combinations of the residues. For a detailed explanation let us consider only the first six positions that were just mentioned. Figure 1-8 shows a schematic of how to generate these ancestral sequences. Four PCR 'Primers' are required to make a full-length product. The first, 'Primer #1', will cover the amino-terminus end and extend through amino acid positions 6 and 8. When this 'Primer' is synthesized it will contain variations that account for the ambiguities at positions 6 and 8. It will contain 50% A and 50% T corresponding to the third position of codon 6 because GAT codes for Asp and GAA codes for Glu, which are the two most probabilistic residues at the position according to PAML, 71.4% and 28.6% respectively. The first position of codon 8 will contain 50% A and 50% T on the corresponding 'Primer' because ACC codes for Thr and TCC codes for Ser. Therefore, 'Primer #1' will actually contain a mixture of four distinct types of primers to satisfy the different combinations of residues, and each primer will constitute 25% of the mixture. 'Primer #2' will have to be synthesized twice. Position 73 contains 3

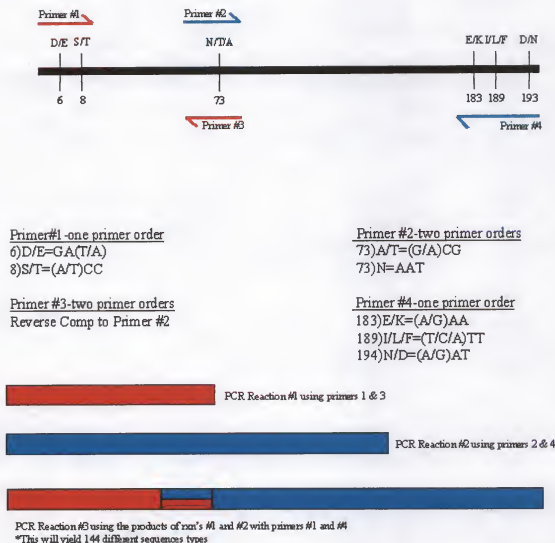


Figure 1-8. Schematic of the PCR reactions that would generate a segment of the *Bacillus* ancestral sequence.

residues that cannot be synthesized by single nucleotide replacements without generating intermediate amino acids. The first reaction will contain nucleotides G and A, each 50%, at the first codon position of residue 73 followed by nucleotides C and G at the second and third position of the codon. The second primer reaction will contain nucleotides A, A and T at the first, second, and third positions, respectively, corresponding to residue 73. 'Primer #2' will then consist of 2-parts reaction one and 1-part reaction two, which will yield 33% of each primer type in 'Primer #2'. 'Primer #3' will be the reverse complement of 'Primer #2'. 'Primer #4' works on the same principle as 'Primer #1'. One primer reaction will generate twelve distinct primer types because there are three ambiguous sites, with one site containing three ambiguities ($2^2 \times 3^1$). Before the full length products are generated, two PCR reactions will be run. The first reaction will use 'Primers #1 and #3'. The second reaction will use 'Primers #2 and #4'. The final reaction will use the products of reactions one and two, overlapping sequence due to 'Primers #2 and #3', along with 'Primers #1 and #4'. Using this method will allow one to generate full-length products that contain every possible combination of the six ambiguous sites (144 different sequences). Although this example does not show all the ambiguous sites in the *Bacillus* ancestral sequence, the same principle can be applied to the other sites until PCR products have been generated that contain all possible combinations of the ambiguous residues over the entire sequence. After all possible sequences have been generated they can be cloned and expressed. A certain percentage of the proteins can be assayed as previously described. Subsequently, a distribution can be generated for the temperature stability of the *Bacillus* ancestral sequence.

The same types of experiments can be performed using the PAML generated *Methanococcus* ancestral sequences. *M. jannaschii* (optimal growth temperature 85°C) and *M. vannielii* (optimal growth temperature 35°C) have been cloned and sequenced (Bult *et al.*, 1996; Lechner and Bock, 1987). The tree topology of EF-Tu archaeal

sequences have been determined (Baldauf *et al.*, 1996). This topology was used as input for PAML. PAML generated the ancestral sequence for the last common ancestor of the *Methanococcus* species. The sequence contains 18 ambiguous sites, 2^{18} . Roughly 260,000 distinct ancestral sequences will be generated using the above procedures. Once a percentage of the sequences are cloned, expressed, and analyzed, another distribution can be created based on temperature stability.

The two distributions created by the *Bacillus* and *Methanococcus* analyses can enable us to compare the distribution generated from EF-Tu ancestral sequences from the base of the bacterial lineage to determine if the ancestor of all bacteria was thermophilic or mesophilic.

Although the experiments in this section were not performed, we believe that they can provide great insight into the question of whether the last common ancestor to all of bacteria was a thermophile or a mesophile. A major limiting factor to the successful completion of these experiments may be the fact that a bacterial tree topology cannot be firmly established. In the next section of Chapter 1 we show how minor branch swapping in a topology can have major effects on the ancestral reconstruction at basal nodes of a tree. In conjunction with the already large numbers of ambiguous residues using the Pace topology, trying to incorporate all of the additional ambiguous residues that alternative tree topologies generate would result in an extremely large number of variants to cover all possible combinations of ambiguous residues. However, we believe that the foundation to these experiments has been firmly established by our initial analyses.

Reconstructing the Divergent Evolution of RNases

Background

Three paralogous lineages of ribonucleases have emerged via gene duplications during the evolution of artiodactyls (camel, cattle, deer, pig, etc.). The pancreatic ribonuclease, RNase A, is one of the best studied of all enzymes (Blackburn and Moore, 1982). This enzyme hydrolyzes the RNA produced from bacteria in the rumen of these animals. The function of brain ribonuclease is currently being examined in the Benner group. Finally, seminal ribonuclease is unique in that it has been hypothesized to play a role in immunosuppressivity whereby the sperm are able to evade the female's immune response during copulation (Soucek *et al.*, 1983). The seminal lineage arose near the time of the divergence of deer from other members of the artiodactyl order, ca. 40 million years ago. Trabesinger-Ruf (1997) demonstrated that bovine seminal RNase could bind to spermatozoa, supporting the hypothesis of this protein's role in immune responses. Subsequently, Raley (2000) attempted to determine when, in the evolutionary history of the seminal lineage, the protein evolved the ability to exert immunosuppressive activity. Based on parsimony analyses, ancestral reconstructions of seminal RNase were generated and tested in the laboratory, Figure 1-9. Two ancestral nodes at the base of the seminal tree, for alternative topologies, and one node at the base of the Bovidae clade were reconstructed. Molecular and paleontological data are unable to unambiguously resolve the relationship of the deer and okapi lineages. These lineages either arose independently ca. 40 million years ago, or together ca. 35 million years ago. The Bovidae arose ca. 5 million years ago. Table 1-1 shows the experimental *in vitro* behaviors of these sequences as compared to extant pancreatic and seminal sequences.

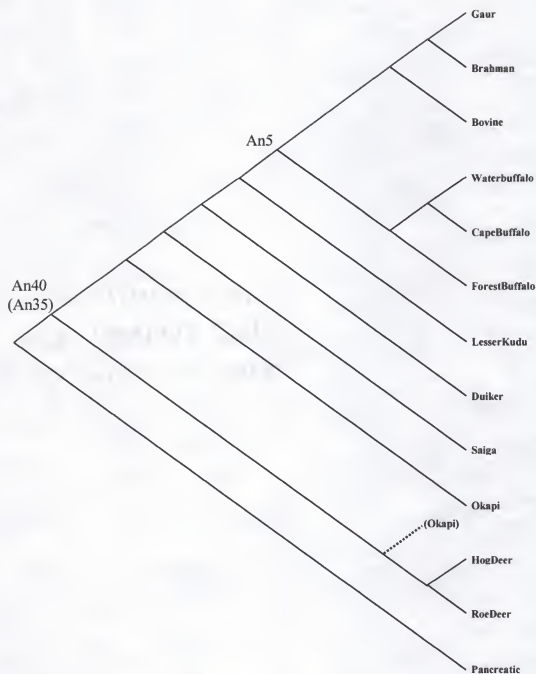


Figure 1-9. Tree topology of seminal RNase with pancreatic RNases as outgroups. This is in agreement with the morphological/paleontological data (Raley, 2000). An40 represents the ancestral sequence at the base of the seminal tree, while An35 represents the ancestral sequence at the base of the tree using the alternative topology of okapi and deer being monophyletic (see dashed line). An5 represents the ancestral sequence of Bovidae.

Some behaviors have not changed significantly during this episode of sequence evolution. This implies that these *in vitro* behaviors are not relevant to the new physiological functions of seminal RNase. In contrast, some *in vitro* behaviors have changed markedly during this episode of possibly-adaptive sequence evolution. The steady increase in immunosuppressivity “up” the tree is culminated by an IC_{50} value of $5\mu\text{g/mL}$ in the extant seminal sequence. Also, the ability to inhibit cell proliferation has constantly increased to $4.2\mu\text{g/mL}$. Therefore immunosuppressivity has only recently arisen through natural selection and can be hypothesized to be the major selective function of seminal RNase according to these assays.

Table 1-1. Comparisons of the *in vitro* properties for ancestral seminal ribonucleases, and the extant bovine seminal and pancreatic ribonucleases. For detailed description of assays see Raley (2000).

Property	Pancreatic	An40	An35	An5	Seminal
k_{cat}/K_m ($\mu\text{M}^{-1}\text{s}^{-1}$)	1.9	2.0	2.2	0.71	1
Poly U	24±4	9.3±2	10±2	12.3±2	14±2
Poly U/A	0.6±0.1	1.6±0.1	3.3±0.2	3±0.2	4.3±0.7
Denature Poly dA/dT	No	No	Yes	Yes	Yes
GM3 Inhibition	No	No	Yes (21%)	Yes (39%)	Yes (45%)
Percentage of swap	<1%	<1%	<5%	30%	70%
IC_{50} MLR $\mu\text{g/mL}$	>100	>100	50±3	12.5±2.6	5±1.2
IC_{50} PHA $\mu\text{g/mL}$	>100	>100	45±2.6	12.5±2.4	4.2±1.3

Maximum Likelihood Models for Reconstructing Ancestral RNases

Reconstructed ancestral sequences using higher order models that include branch lengths and substitution matrices can result in different reconstructions of residues than simpler parsimony-based analyses. In Figure 1-10, we show how different likelihood models can reconstruct different residues at some positions. This figure shows how incorporating different models can lead varying results and how each model can calculate a high probability associated with their results, regardless of whether the result is “correct”. Thus, caution must be taken when incorporating models of sequence evolution into molecular analyses.

It is unclear whether the gene duplication event that gave rise to seminal RNase resulted from a duplication of the pancreatic or brain ribonuclease gene (although some phylogenetic analyses suggest pancreatic as the precursor). Since the composition of outgroups can effect reconstructions at the base of the ingroup tree, we tested for these effects on the seminal tree, part (a). As expected, reconstructions were only effected at the base of the tree (An40 and An35). The reconstructions at these nodes are clearly influenced by the sequences leading-to and leading-away from these nodes. In addition, we tested the reconstructions using an amino acid model and a codon model. The amino acid model relies on a replacement matrix based on observed/expected values for a large number of analyzed replacements (e.g., JTT and Dayhoff matrices). The codon model relies directly on the data since the substitution matrix is formulated from the input data set. Both models have their advantages and disadvantages. The codon model works from a 64-by-64 matrix (minus the stop codons) but can over-parameterize the analysis,

whereas the amino acid model works from a 20-by-20 matrix but may not be complex enough to accurately represent the mode of evolution for certain data sets.

Differences between the two models, and between the two outgroups, are apparent for all three ancestral nodes. For example, position 22 has a serine predicted by the amino acid model for An40 with the pancreatic outgroup, but asparagine is predicted as the most likely ancestral residue by the codon model for the same node and outgroup. The effect of outgroup composition is seen at position 65 for An40. A glutamine is predicted with a high degree of probability by both the amino acid and codon models using pancreatic RNases as the outgroup, although lysine is predicted by both models with brain as the outgroup. A number of examples can be highlighted. It is thus important to understand the effect of incorporating different models into ancestral reconstruction analyses.

Figure 1-10. Comparing seminal RNase ancestral reconstructions between parsimony and various models of maximum likelihood. All maximum likelihood analyses were performed using PAML, under conditions as stated in the text.

a) The extant amino acid RNase *A* (pancreatic) sequence is listed in single letter code. Ancestral residues are listed when the reconstruction differs from the extant sequence. An40, An35 and An5 seminal reconstructions are based on parsimony analyses with the complete pancreatic clade as the outgroup. Maximum likelihood reconstructions are listed below the parsimony reconstructions for the appropriate nodes using a subset of pancreatic or brain sequences for the outgroup. The posterior probabilities using a codon or amino acid model are listed to the left and right of the reconstruction, respectively, only when the reconstruction differs from the parsimony reconstruction. When room is not available, the probability using the codon model is listed to the right in parentheses. Also, the extant bovine seminal residues are listed at positions that differ from the extant pancreatic sequence; b) Same format as in part (a). However, only the reconstructions corresponding to node An40, and using pancreatic sequences as the outgroup, are listed. This figure attempts to elucidate the influence of outgroup sample size and 1x4 versus 3x4 base frequency tables, under the codon-based analysis, on ancestral reconstructions.

(a)	10	20	30	40	50
<i>RNase A</i>	KETAAAKFER	QHMDSSSTGAA	SSSNYCQMM	KSRNLTKDRC	KPVNTFVHES
An40	S	GS PS	L	FC KM QGK	
Pancr outgroup		91S82 17S85 09P17 83N15			
Brain outgroup	02E47 35R37 21W11 46G01	71S68 99N99 28P32			
An35	S	GS PS	L	FC KM QGK	
Pancr outgroup			92N95		
Brain outgroup	04E60 39R38 56G00	99N100			
An5	S	GS PS N	L	FC KM QGK	
Pan/B outgroup					
<i>BS-RNase</i>	S	GN PS	L	CC KM QGK	

	60	70	80	90	100
<i>RNase A</i>	LADVQAVCSQ	KNVACKNGQT	NCYQSYSTMS	ITDCRETGSS	KYPNCAYKTT
An40	Q	K A	N A H		
Pancr outgroup	91D80Q97(98) 09N20	08A78 92T21	98S89		
Brain outgroup	52N95K98(99) 48D05	98T99	87S82		
An35	K	K A	N A H		
Pancr outgroup	98Q97	95T92	44T61 56S39		
Brain outgroup	50N75K97(99) 50D25	99T99	81S72 16T26	60H85 40R14	
An5	K	K T	K T R		
Pan/B outgroup					
<i>BS-RNase</i>	K	K T	K T R		

	110	120	124
<i>RNase A</i>	QANKHIIVAC	EGNPFYVPVHF	DASV
An40	N	N Y	
Pancr outgroup	59Q80E99(98) 47R20	58N96 41K	
Brain outgroup	48Q40E99(98) 52R40	38N80 61K20	
An35	N	K Y	
Pancr outgroup	E99(98)	59N96 41K	
Brain outgroup	E99(98)	44N83 55K17	
An5	VE R	A K Y	
Pan/B outgroup	70H85 36R15	11A64 76E17 13G18	
<i>BS-RNase</i>	VE	G K S	

Figure 1-10 continued

(b)	10	20	30	40	50
<i>RNase A</i>	KETAAAKFER	QHMDSSSTSAA	SSSNVCNQMM	KSRNLTKDRC	KPVNTFVHES
An40	S	GS PS	L	FC KM QGK	
Pancr outgroup (partial 3X4)		91S82 17S85			
Pancr outgroup (partial 1X4)		09P17 83N15			
Pancr outgroup (complete 3X4)		79S 80S			
<i>BS-RNase</i>	S	GN PS	L	CC KM QGK	

	60	70	80	90	100
<i>RNase A</i>	LADVQAVCSQ	KNVACKNGQT	NCYQSYSTMS	ITDCRETGSS	KYPNCAYKTT
An40	Q	K A	N A H		
Pancr outgroup (partial 3X4)	91D86Q97(98)	08A78	98S89		
Pancr outgroup (partial 1X4)	09N20	92T21			
Pancr outgroup (complete 3X4)	79D Q--(95)	60A	99S		
<i>BS-RNase</i>	K	K T	K T R		

	110	120	124
<i>RNase A</i>	QANKHIIVAC	EGNPYVPVHF	DASV
An40	N	N Y	
Pancr outgroup (partial 3X4)	57Q80E98(98)	58N96	
Pancr outgroup (partial 1X4)	47R20	41K	
Pancr outgroup (complete 3X4)	79Q E--(99)	94N	
<i>BS-RNase</i>	VE	G K S	

The original parsimony-based analysis using amino acid data, and pancreatic RNases as the outgroup, identified two positions that differed between nodes An40 and An35 (positions 65 and 113). Our maximum likelihood analysis may have “resolved” these differences, as they no longer exist under the amino acid model with the pancreatic outgroup. However, four additional positions are now highlighted in our analyses comparing An40 and An35 for amino acid data with the pancreatic outgroup (positions 19, 22, 64 and 70)

The effects of incorporating base frequency tables into the codon model were tested for An40 using the complete pancreatic outgroup or a sample of the sequences from the outgroup. Table 1-2 shows the unequal distribution of base frequencies for the first, second and third codon positions of seminal ribonuclease. A 3x4 base frequency table incorporates this non-uniform GC distribution, whereas a 1x4 table only incorporates the base frequencies treating the codon as a single unit. Positions 22, 64 and 113 are influenced by the incorporation of these two different parameters, Figure 1-10 part (b). Interestingly, only sites 22 and 113 are influenced when using either the complete or partial pancreatic outgroup. Parameters, models and data input can all affect the outcome of ancestral reconstructions. It is therefore necessary to incorporate various models into the computational reconstruction analysis before any sequences are generated in the laboratory.

Table 1-2. Base frequencies used in the codon model analysis for ancestral reconstructions of seminal ribonuclease. The 1x4 table only includes the mean base frequencies for the codons. The 3x4 table includes the base frequencies for all three codon positions when reconstructing the ancestral sequences. Therefore, only by incorporating the 3x4 table can the extreme GC bias of the third codon position be considered.

Codon position	T	C	A	G
1	22%	17%	34%	27%
2	19%	25%	36%	20%
3	13%	47%	10%	30%
Mean	18%	30%	27%	26%

An Alternative View of Reconstructing Ancestral Sequence "Space"

Although the methods of reconstructing ancestral sequences have greatly improved in the past few years, the outputs of these methods have not been comprehensively studied. This study is unique in that it marks the first time a data set has been analyzed by multiple models and multiple parameters with the intention of actually reconstructing these ancestral sequences in the laboratory. Generating reconstructed sequences has always relied on individual nodes within the tree topology. However, our analysis strongly suggests that it may not be possible to reconstruct these points on a tree. We would, therefore, like to advocate a new approach for reconstructing ancestral sequences along a tree. This is conceptually shown in Figure 1-11. It is possible that we may never know whether deer and okapi are monophyletic, or whether the pancreatic or brain outgroup duplicated to give rise to the seminal lineage. However, we should not be paralyzed by this lack of information. By incorporating historical and evolutionary ambiguities, we have shown that it is possible to calculate the ancestral sequence

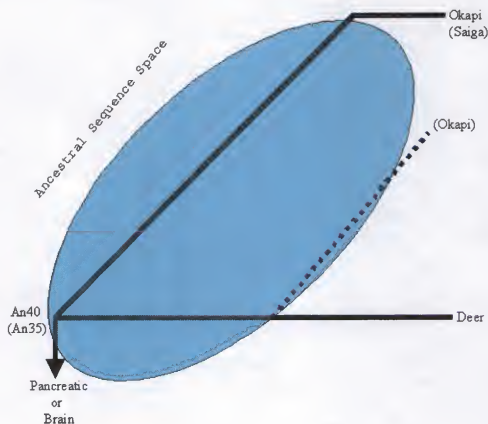


Figure 1-11. Representation of the mutation space that ancestral seminal RNase sequences have evolved through. It is important to note that this space is not random space, but rather mutation space in which natural selection has presumably eliminated any deleterious mutations. The dashed line represents the alternative tree topology with corresponding nodes in parentheses. Reconstructing ancestral sequences from this space allows one to identify all possible mutations regardless of the exact point of the mutation on the tree, and is therefore advantageous over other methods that reconstruct sequences at individual points on the tree.

population around an unknown point in evolution. This has clear advantages since the analysis does not depend on, for example, what the true outgroup is. Also, we can correlate rapid sequence evolution with change in function along the tree. Obviously, interpretation of the behavioral properties of the reconstructed proteins will depend on these ambiguities. In short, we believe this method will prove valuable for understanding molecular evolution within a paleogenomics framework and assist in understanding the mechanisms by which divergent protein behavior evolves.

CHAPTER 2

DETECTING FUNCTIONAL DIVERGENCE IN BIOLOGICAL SEQUENCES: HISTOGRAM APPROACH USING ORIGINAL RATE DIFFERENCES

Background

The divergent evolution of protein sequences from genomic databases can be analyzed using different mathematical models. The most common treat all sites in a protein sequence as equally variable. More sophisticated models acknowledge the fact that purifying selection generally tolerates variable amounts of amino acid replacement at different positions in a protein sequence. In their "stationary" versions, such models assume that the replacement rate at individual positions remains constant throughout evolutionary history. "Non-stationary" covarion versions, however, allow the replacement rate at a position to vary in different branches of the evolutionary tree. Recently, statistical methods have been developed that highlight this type of variation in replacement rates. Here, we show how positions that have variable rates of divergence in different regions of a tree ("covarion behavior"), coupled with analyses of experimental three-dimensional structures, can provide experimentally testable hypotheses that relate individual amino acid residues to specific functional differences in those branches. We illustrate this in the elongation factor family of proteins as a paradigm for applications of this type of analysis in functional genomics generally.

Elongation factors Tu (EF-Tu) and 1α (EF- 1α) are homologous proteins essential to translation in bacteria and eukaryotes, respectively (Krab and Parmeggiani, 1998; Negrutskii and El'skaya, 1998). These GTPases catalyze the binding of aminoacyl-

transfer RNAs (aa-tRNA) to the A-site of the ribosome. As they are among the slowest evolving proteins known, EFs are commonly used to study cellular functions (Negrutskii and El'skaya, 1998; Yang *et al.*, 1990; Duttaroy *et al.*, 1998) and to root the universal tree of life (Lopez *et al.*, 1999; Baldauf *et al.*, 1996). This sequence stability presumably reflects enormous functional constraints on the divergent evolution of EFs, highlighting their central role in translation since the last common ancestor of the three primary domains of life (Benner *et al.*, 1989). Nevertheless, EF-Tu and EF-1 α differ in several of their specific functions (Krab and Parmeggiani, 1998; Negrutskii and El'skaya, 1998). For example, bacterial EF-Tu binds GDP ~100 fold tighter than GTP. Eukaryotic EF-1 α , in contrast, binds both with similar affinities. EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts. EF-1 α requires the multi-subunit nucleotide exchange factor EF-1 $\beta\gamma\delta$. EF-1 α also interacts with the eukaryotic cytoskeleton and may thereby play a role in cellular transformation and apoptosis (Negrutskii and El'skaya, 1998; Yang *et al.*, 1990). EF-Tu can have no such role in bacteria.

These shifts in function must correspond at some level to changes in protein sequence. Thus, functional changes can leave signatures in the sequences of a protein family, which can then be detected with a well constructed history of their relationships and replacements. In many cases, it appears possible to identify this record from the background noise of molecular evolution. In alcohol dehydrogenase (Benner *et al.*, 1998) and superoxide dismutase (Miyamoto and Fitch, 1995), for example, previous studies have shown that variable replacement rates at specific positions can generate inferences relating changes in sequence structure to those in function. These proteins, however, have diverged far more rapidly than EFs. Further, these studies have used neither the full

power of a mathematical evolutionary (Benner *et al.*, 1998) nor crystallographic (Miyamoto and Fitch, 1995) analysis. We show here how this combination is of value in functional genomics, even in proteins not generally regarded as good examples of functional divergence.

From a mathematical perspective, the most common way to model rate heterogeneity among sequence positions is the gamma distribution, with its shape parameter alpha (α) (Swofford *et al.*, 1996; Yang, 1996). This distribution can accommodate a wide range of rapidly and slowly evolving sites. However, this model assumes a stationary substitution process, whereby positions retain their same relative rates of change throughout evolutionary history. This assumption is not expected to hold entirely true for proteins that change function. As an alternative, the covarion model proposes that the replacement rates of amino acid positions can change over time (Miyamoto and Fitch, 1995; Fitch and Markowitz, 1970; Tuffley and Steel, 1998; Gu, 1999; Morozov *et al.*, 2000). Although EFs might be expected to follow only a gamma model given their overall functional conservation, previous studies have instead suggested that a covarion process is needed to adequately describe their evolution (Lopez *et al.*, 1999; Lockhart *et al.*, 1998; Moreira *et al.*, 1999). This conclusion is examined more closely in this chapter and forms the basis of our integrated evolutionary and structural biology analyses of functional divergence between EF-Tu and EF-1 α .

Methods

Thirty EF sequences were aligned by DARWIN (Benner, 1998) and then modified according to the secondary structures of EF-Tu for *Escherichia coli* (PBD accession number 1EFC) (Song *et al.*, 1999) and *Thermus aquaticus* (PBD accession number 1TTT) (Nissen *et al.*, 1995). This approach resulted in a multiple sequence alignment

(MSA) with 380 aligned positions (cf. Moreira *et al.*, 1999). Maximum likelihood (ML) estimations of α and the replacement rates per site for all 380 aligned positions of EF-Tu versus EF-1 α were accomplished with PAML, v2.0, and its implementation of the Jones, Taylor, and Thorton model, with rate heterogeneity among sites according to the gamma distribution (JTT- Γ) (Yang, 1997). The Proportional, Poisson, and Dayhoff models for protein sequences were rejected as less appropriate for EFs on the basis of their log-likelihood ratio tests (Huelsenbeck and Rannala, 1997). The phylogeny in these ML analyses followed that of Bauldauf *et al.* (Bauldauf *et al.*, 1996), except for the topological positions of *Chlorobium* and *Salmonella*. As Bauldauf *et al.* did not consider these two species, their topological positions were based on our follow-up ML analyses with MOLPHY, v.2.3 (Adachi and Hasegawa, 1997).

Parametric bootstrapping (evolutionary simulations) was conducted with PAML to calculate the standard deviations (SD) of the α estimates for bacteria alone, eukaryotes alone, and both groups combined (Huelsenbeck *et al.*, 1995). These simulations (20 per group) relied on the accepted tree and subtrees of bacteria and eukaryotes, their ML estimates of branch lengths and α , and the JTT- Γ model. In turn, subsampling experiments with bacteria alone, eukaryotes alone, and the two groups combined were completed to test for sample-size effects on their estimations of α (Sullivan *et al.*, 1999). In these experiments, 20 random subsets apiece were generated for all odd-numbered subsamples from 5 to 11, 13, and 27 for bacteria, eukaryotes, and both groups, respectively. The α parameter was then re-estimated for each random subsample using the same ML conditions as before. In recognition of their greater numbers, the

subsampling trials with both groups combined were stratified such that an extra eukaryotic sequence was selected relative to bacteria.

Normal distributions, sample kurtosis, skewness, and normality tests were all determined with SAS/Graph, rel. 6.03 (SAS, 1988). Visualization of protein structures was accomplished with Chemscape Chime, rel. 2.0.3 (www.mdli.com) and Protein Explorer, rel. 1.46 (www.umass.edu/microbio/chime/explorer).

Covarian Analyses, Structural Biology, and Hypothesis Generation

The results of the log-likelihood tests are shown in Table 2-1. No significant differences in log-likelihood scores were observed between the Dayhoff-gamma, Dayhoff-f-gamma, and JTT-gamma models. The score obtained by integrating the JTT-gamma-rho was significantly different than the JTT-gamma. Rho (ρ) models the extent to which adjacent sites are evolving non-independently (i.e., correlation in evolutionary rates for neighboring sites) and varies between 0 (adjacent sites are evolving independently) and 1 (every site's evolutionary rate is correlated with its nearest neighbor's rate) (Yang, 1995). To test whether incorporating these various models had any effect on the estimation of replacement rates, we plotted the rates comparing JTT-gamma versus JTT-gamma-rho (JTT- Γ - ρ). As seen in Figure 2-1, there is little or no difference in estimated rates for both bacteria and eukaryotes when using either model. The correlation coefficients were 0.99508 and 0.99579 for bacteria and eukaryotes, respectively. The near identity of the rate estimates for the JTT-gamma and JTT-gamma-rho alternatives support the conclusion that such ML calculations are robust to the chosen model of sequence evolution (Yang, 1995). However, the significantly better fit of the observed data to the JTT-gamma-rho model (with its ML estimate of $\rho = 0.466$) argues

for a relatively strong positive correlation between the individual rates of adjacent sites.

In particular, this potential lack of independence is of special interest to the covarion sites, because of their evident structural and functional ties (see below).

Table 2-1. Log-likelihood scores for the EF data set using different models of amino acid evolution. Log-likelihood Ratio Test (LRT) is a measure of the significance between two competing nested-models. In this case, there is one degree of freedom for the test when comparing the JTT-gamma and JTT-gamma-rho models $\{\delta=2[-9287-(-9307)]=40\}$.

Model	Log-likelihood score
1) Poisson	-10741
2) Proportional	-10504
3) JTT-f	-9779
4) JTT	-9752
5) Dayhoff-f	-9748
6) Dayhoff	-9744
7) JTT-f-gamma	-9326
8) JTT-gamma	-9307
9) Dayhoff-f-gamma	-9302
10) Dayhoff-gamma	-9299
11) JTT-gamma-rho	-9287
LRT between 8 & 11	P<0.005

Our ML analyses of EF-Tu and EF-1 α revealed a non-stationary α for different regions of the tree (Figure 2-2). An α of 0.78 was calculated for the entire tree, with a SD of 0.05 from parametric bootstrapping. In contrast, the α values for both the bacterial and eukaryotic subtrees were significantly lower [$\alpha = 0.46$ (0.04) and $\alpha = 0.38$ (0.04), respectively]. Thus, a more uniform distribution of rates among sites was suggested when the two groups were considered together, rather than separately. Gu (1999) statistically proved that such an increase in α is expected when the variable positions of one group are not the same as those of another (i.e., when the sequences are evolving under a non-stationary covarion process). For a schematic representation of this concept

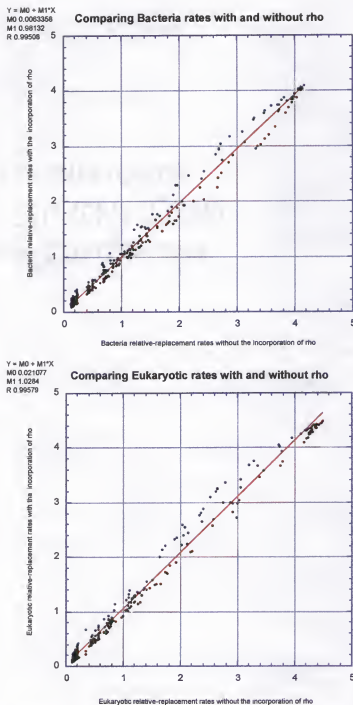


Figure 2-1. The effects of incorporating the rho parameter when estimating bacteria and eukaryotic relative-replacement rates. The rho parameter signifies whether adjacent sites (i and $i+1$) are evolving independently or dependently (correlated).

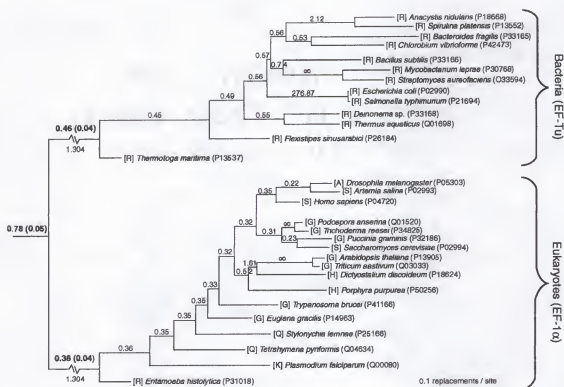


Figure 2-2. Accepted phylogeny for bacteria and eukaryotes used in the ML analyses of their EF-Tu and EF-1 α sequences. These sequences are from SWISS-PROT, with their accession numbers given in parentheses next to their species. Brackets refer to the amino acids of the two groups at position 305, a site illustrating a covarion pattern of sequence conservation in bacteria but considerable variation in eukaryotes. Branch lengths of this tree are drawn proportional to their ML estimates, except for the two longest internodes leading to bacteria and eukaryotes (both 1.30 replacements/site). Total tree length is 7.34 replacements/site (2.54 and 2.37 replacements/site for bacteria and eukaryotes alone, respectively). Numbers above internal branches represent the ML estimates of α for the corresponding group or subgroup of bacteria and/or eukaryotes. Standard deviations, as calculated from twenty rounds of parametric bootstrapping, are given in parentheses for the α values of bacteria, eukaryotes, and the two groups combined.

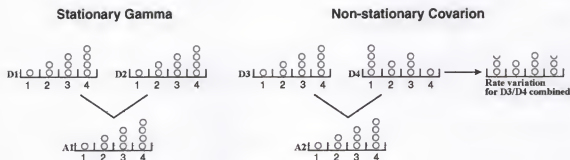


Figure 2-3. A schematic representing the differences between the stationary gamma and non-stationary covarion (gamma) models. These two examples both begin with an identical ancestral population of sequences (A1 and A2, respectively). This ancestral population contains the same site-by-site rate variation in both examples as represented by the number of circles in each site; the first site is slowly evolving, sites 2 and 3 are moderately evolving, while site 4 is rapidly evolving. In the case of stationary gamma evolution, both descendent populations D1 and D2 contain the same numbers of slowly, moderately and rapidly evolving positions as their ancestor. Thus, their rate heterogeneity of sites is the same and the α value for each is expected to be the same.

The opposite condition holds true for the covarion example. Although the descendent populations D3 and D4 contain the same numbers of slowly, moderately and rapidly evolving positions as their ancestral state A2, the identities of these sites have changed in D4. Site 1 is now rapidly evolving whereas site 4 is slowly evolving. It is this non-stationary behavior that distinguishes the covarion and gamma models. This type of covarion behavior causes spurious estimates of the gamma distribution. The estimated alpha values of populations D3 and D4 are the same as A2 when analyzed individually since each descendant has one slow, one rapid, and two moderate sites. However, when D3 and D4 are combined, the estimated alpha value increases due to an averaging effect of the rate variation. Thus, as statistically proven by Gu (99), an increased α value for combined data is one sign of a covarion process.

see Figure 2-3. The rate profile of the replacement rate differences shows that the individual site rates are not uniform between the bacterial and eukaryotic lineages (Figure 2-4)

The distribution of rate differences per site between bacterial and eukaryotic EFs was leptokurtotic; i.e., over- and under-represented in the mean and tails versus “shoulders,” respectively, relative to the expectations of a normal distribution (Figure 2-5). Nearly 50% of the positions had essentially the same rate in the two groups (rate differences of <0.5 replacements/site/unit evolutionary distance), as expected under a stationary gamma process. However, 17 sites were evolving >2 SD faster in bacteria than eukaryotes, while 19 were changing >2 SD faster in eukaryotes than bacteria (Figure 2-5). These sites representing 10% of the MSA are suggestive of a covarian process in the EF-Tu/EF-1 α family.

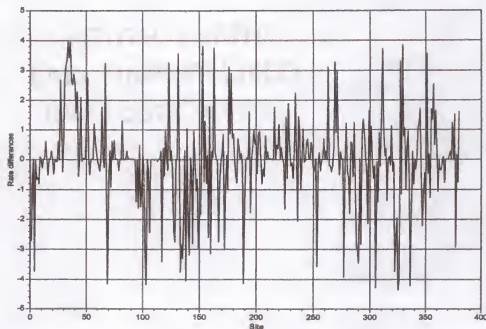
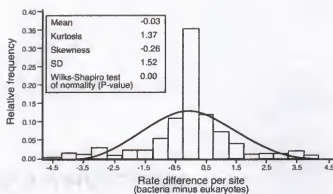


Figure 2-4. Rate profile for the replacement rate differences between the bacteria and eukaryotic lineages. Peaks at positive and negative values indicate a non-stationary process for the evolution of rates in EFs

By integrating structural data with these ML rate differences, this initial pool of 36 sites can be further reduced to a subset of those positions that are most likely involved in the functional shifts between EF-Tu and EF-1 α . For example, ten sites in and around the region binding tRNAs are evolving >2 SD faster in either bacteria or eukaryotes (Figures 2-5 and 2-6). These rate changes can be correlated to a difference in biochemical function between EF-Tu and EF-1 α . EF-1 α /GDP binds charged and uncharged tRNAs, whereas EF-Tu/GDP does not. Crystallographic data for EF-Tu reveals a major conformational shift between the GDP- and GTP-bound states, whereby the tRNA-binding site of the former is disrupted (Figure 2-6). In contrast, available data for EF-1 α suggest that this conformational shift does not occur. This correlation between rate differences and protein structure/function leads to the hypothesis that at least some of these ten positions are responsible for the different interactions of EF-Tu and EF-1 α with tRNA. This hypothesis can now be tested by introducing into EF-1 α the residues of EF-Tu at these positions (Golding and Dean, 1998). The prediction is that these introductions will result in a variant of EF-1 α /GDP that does not bind uncharged tRNA.

Similarly, eight sites in and around the region where nucleotide exchange factors bind are evolving >2 SD faster in eukaryotes than in bacteria (Figures 2-5 and 2-6). EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts, whereas EF-1 α depends on the multi-subunit EF-1 $\beta\gamma\delta$. The rate differences for these eight sites lead to the hypothesis that the surface area of EF-1 α in contact with its nucleotide exchange complex is different than that for EF-Tu. This difference is consistent with the divergent structures of their respective nucleotide exchange factors (Krab and Parmeggiani, 1998; Negrutskii and El'skaya, 1998).



Eukaryotes>bacteria (i.e., left tail)	Secondary structure	Properties, function
4, 69, 160, 290	β , β , L, β	Surface, no known function
117	α	Surface, in proximity to EF-Ts and kirromycin binding
102-103, 133, 135, 138, 141, 336	L, α , L α , α , L	Surface, all residues bind EF-Ts
189	L	On loop connecting domains 1 and 2
325-326	β	Surface, 2-3 residues away from aa-tRNA binding
253, 277, 305, 322	L, L, L, β	Surface, all residues bind aa-tRNA

Bacteria>eukaryotes (i.e., right tail)	Secondary structure	Properties, function
32-36	L	Surface, possible localization sites or ribosome binding
131	α	310 Helix binds GTP/GDP, faces away from nucleotides
153, 163	α , β	Interior, no known function
269	α	310 Helix, in proximity to aa-tRNA binding
263, 327, 329	β , β , β	Surface, 3-4 residues away from aa-tRNA binding sites
67, 123, 176, 311, 351	L, L, α , L, L	Surface, possible localization sites

Figure 2-5. Rate differences per site between bacteria and eukaryotes. Top part of figure is the histogram of the site-by-site rate differences for the 380 aligned positions of bacteria minus eukaryotes. Sample kurtosis and skewness measure the “peakedness” and asymmetry of the histogram relative to the superimposed normal distribution, respectively. Bottom part of figure represents the amino acid positions in the left and right tails of the histogram (i.e., those with rate differences of >2 SD between the two groups). Numbering refers to positions in the MSA. “ α ”, “ β ”, and “L” refer to α -helices, β -strands, and loops, respectively, following the three-dimensional structure of EF-Tu (Figure 2-6).

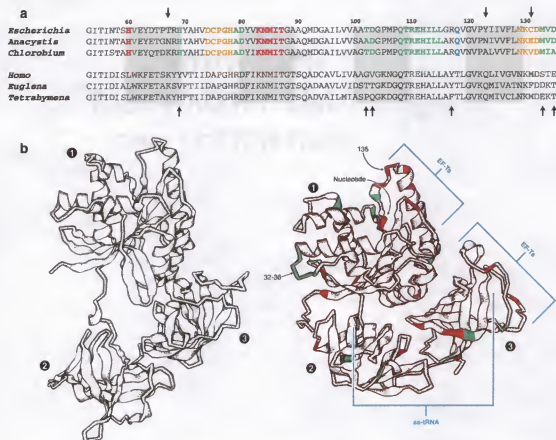


Figure 2-6. MSA for EFs and tertiary structures for EF-Tu. (A) MSA for the ligand-binding region at the NH₂-terminus of three representative bacteria and three eukaryotes (top and bottom, respectively). This MSA highlights the key residues for aa-tRNA (red), EF-Ts (green), and nucleotide (yellow) binding and for kirromycin resistance (cyan), as determined for bacterial EF-Tu. Arrows, above and below the MSA, correspond to those sites that are evolving >2 SD faster in bacteria than eukaryotes, and vice versa, respectively (positions 67, 69, 102-103, 117, 123, 131, 133 and 135) (Figure 2-5). (B) Tertiary structures of the GDP- and GTP-bound states for EF-Tu from *E. coli* and *T. aquaticus*, respectively. Here, green and red in the GTP confirmation highlight those sites that are evolving >2 SD faster in bacteria than eukaryotes, and vice versa, respectively (Figure 2-5).

Perhaps the most intriguing functional difference between the two EFs is the ability of EF-1 α to bind to actin, the main component of the eukaryotic cytoskeleton. This function, together with the ability of EF-1 α (but not EF-Tu) to bind to uncharged tRNAs, may be important as a mechanism for tRNA channeling from the ribosome back to the nucleus (Negruksii and El'skaya, 1998; Grosshans *et al.*, 2000). Bacteria, of course, do not require channeling, thereby obviating the need for binding of uncharged tRNAs by either the GDP- or GTP-states of EF-Tu. Relatively rapid sequence evolution is a general characteristic of surface residues that are not involved in protein-ligand interactions (Lichtarge *et al.*, 1996). Nine surface residues to which other contacts cannot be definitively assigned from biochemical and structural data were evolving >2 SD faster in EF-Tu than EF-1 α (Figures 2-5 and 2-6). These rate differences suggest the hypothesis that at least some of these residues in EF-1 α are in contact with the actin cytoskeleton.

Positions 32-36 are conserved in EF-1 α , but variable in EF-Tu (Figures 2-5 and 2-6). In EF-Tu, biochemical and three-dimensional structural data show that this region is in proximity to the ribosome (Peter *et al.*, 1990; Ban *et al.*, 1999). In EF-1 α , positions 32-36 are followed by an insertion that is suggestive of a binding site with its characteristic charged amino acids and hydrophobic residues. In combination with its conserved residues 32-36, this insertion is predicted to introduce a regular secondary structural element of an α -helix (Benner *et al.*, 1998; Rost and Sander, 1993) that may reflect a difference in ribosomal structure and binding between bacteria and eukaryotes. Thus, another testable hypothesis is suggested by the integration of rate differences with protein structure and function.

How robust are our hypotheses with respect to the current sample of sequences? This question follows from the recent demonstration by Sullivan *et al.* (1999) that ML

estimates of rate variation among sites may be sensitive to taxon sampling. In our subsampling experiments, estimates of α were found to be upwardly biased for the smaller samples of all three groups (Figure 2-7). Nevertheless, the same major difference between bacteria and eukaryotes alone versus combined was evident, regardless of the sample size. Also, α remained largely unchanged (within the range of statistical error) with the inclusion of 40 and 15 additional sequences from SWISS-PROT for bacteria (0.48) and eukaryotes (0.35), respectively. Given our initial focus on the fluctuating estimates of α for bacteria and eukaryotes, our study did not consider Archaea. However, our more recent investigations of EFs document that this group is defined by an α (0.88) that is more similar to the combined estimate for bacteria and eukaryotes than to their separate values. Collectively, these various results argue against sampling error as an explanation for the non-stationary behavior of α for EF-Tu versus EF-1 α .

We also tested the effect of estimating replacement rates when using different alpha values. Two sets of rates were estimated for bacteria-only using alpha equal to 0.46 and 0.78 (correct and incorrect values, respectively). The same was done for eukaryotes-only with alpha equal to 0.38 and 0.78. A concatenated data set of bacterial rates as estimated with alpha=0.46 and eukaryotic rates as estimated with alpha=0.38 was created. Another concatenated data set was created as above, but included rates as estimated with alpha=0.78 for both lineages. Figure 2-8 shows the results of this comparison. Although the correlation coefficient indicates that there is no difference in estimated rates when using either the correct or incorrect alpha value, a subtle but significant difference can be seen upon closer inspection. The slope of the linear fit is 0.75843 and the y-intercept is 0.17692. These results demonstrate a clear bias when using different alpha values. The

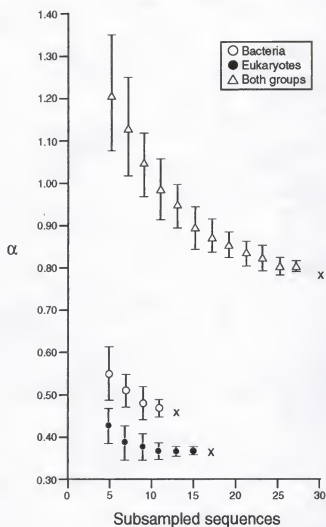


Figure 2-7. The effect of sequence sample size on the ML estimation of α for bacteria alone, eukaryotes alone, and the two groups combined. "X"s correspond to the final estimates of α for each group. Twenty subsampling experiments were completed for each sample size of a group, with the results summarized as means and their SD.

$Y = M0 + M1 \cdot X$
 $M0: 0.17692$
 $M1: 0.75843$
 $R: 0.99517$

All 760 sites (380 bac + 380 eukary) comparing the correct alpha values vs. "incorrect" values

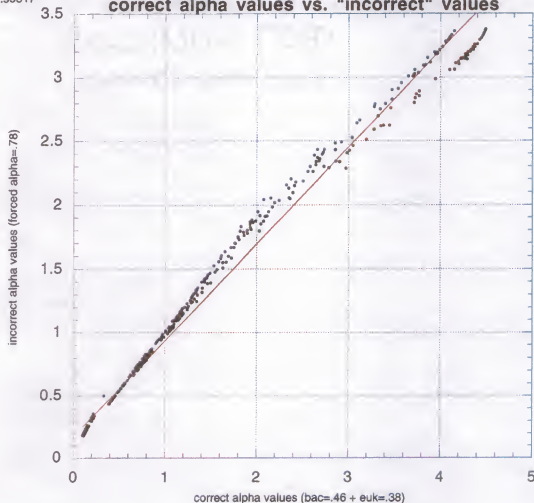


Figure 2-8. Comparing the bacterial and eukaryotic replacement rate estimates when using the alpha values of 0.46 and 0.38 from the individual lineages, respectively, versus the alpha value of 0.78 from the complete tree. The values on the x- and y-axes represent the relative-replacement rates (replacements/site/unit evolutionary time)

incorrect alpha value of 0.78 forces the replacement rates to have a more uniform distribution, and since most of the sites have low replacement rates, the incorrect alpha doesn't allow truly rapidly evolving sites to have a high rate. In addition, notice how the y-coordinate peaks at a value of ~ 3.4 , while the x-coordinate peaks at ~ 4.2 . These two peaks would be nearly identical if alpha values had no effect on the estimates of replacement rates. Thus, we have again demonstrated that replacement rates are not stationary for elongation factors and that rates are intimately correlated to alpha, as expected.

Covarian Approaches and Functional Genomics

Functional genomics is the bridge between computational and experimental biology (Bork and Koonin, 1998; Benner *et al.*, 2000). The field combines sequence data with general knowledge to generate testable hypotheses about the biological functions of genes and proteins. Today, most hypotheses in the field are generated from sequence similarity searches with BLAST (Lipman and Pearson, 1985) or FASTA (Altschul *et al.*, 1997). The function of the probe sequence is assumed to equal that of the best annotated hit that is recovered in these similarity searches.

Functional genomics is actively seeking tools to detect changes in protein function from their sequences and estimated history (Gu, 1999; Yang *et al.*, 2000). The best-known approach for this purpose uses the ratio of nonsynonymous to synonymous substitutions to identify potential cases of functional change (Yang *et al.*, 2000; Li *et al.*, 1985; Messier and Stewart, 1997). This approach, however, suffers as a signature of functional change among distant branches, since silent sites quickly lose their signal as they become saturated with substitutions. Shifts in protein function can also be deduced

from instances of convergent or parallel evolution (Messier and Stewart, 1997). In turn, functional constraints can be detected as compensatory covariation, whereby different residues in contact are sequentially replaced in a way that conserves some overall physical property (Chelvanayagam *et al.*, 1997).

The covarion approach now offers another tool for studying the evolution of protein function (Gu, 1999). Variability is a feature of a position which reflects its relation to selected function. Thus, changes across groups in the variability of their sites offer insights into which positions of a protein may be most responsible for its functional shifts. If the variability of many positions changes, then the inference can be made that the protein has acquired a new function (or lost its function). However, this study with EFs illustrates how much our concept of function is contingent on one's perspective and how subtle such shifts can be. In detail, EF-Tu and EF-1 α function in different ways, even though their overall role in translation has remained the same. These more subtle, but nevertheless, significant functional differences involve on the order of 10% of the sites according to our covarion analysis (Figure 2-5).

Our approach integrates structural data with a covarion-based evolutionary analysis to improve the identification of those relatively few sites that are largely responsible for the functional differences between EF-Tu and EF-1 α . Together, these two sources of information allow us to target specific positions and residues for the direct experimentation of their effects on the function of EFs. Of particular interest are the surface residues that are evolving >2 SD slower in eukaryotes than in bacteria. If confirmed by direct testing, the involvement of at least some of these sites in binding EF-1 α to actin would constitute one of the only examples where metabolic channeling, long an issue in central pathways, has left a signature in the sequences themselves (Reddy and

Pardee, 1980). It is as a tool for hypothesis generation and experimental design that covarion-based evolutionary studies, coupled with structural biology, will make their greatest contributions to functional genomics.

CHAPTER 3

DETECTING FUNCTIONAL DIVERGENCE IN BIOLOGICAL SEQUENCES: HISTOGRAM APPROACH USING LOG-TRANSFORMED RATE DIFFERENCES

Background

In the previous chapter we described a new approach for the identification of sequence positions with changing evolutionary rates. This approach was based on the non-stationary covarion model, whereby a site can be rapidly evolving in one group but highly conserved in another (Fitch and Markowitz, 1970; Miyamoto and Fitch, 1995). This covarion-based approach relied on maximum likelihood (ML) methods to estimate the evolutionary rates of sites in one group, then in another. Those sites with the greatest changes between groups were identified from a frequency histogram of their rate differences. As an illustration of this approach, we provided a detailed ML analysis of elongation factors for bacteria versus eukaryotes (380 aligned positions for 13 EF-Tu versus 17 EF-1 α sequences, respectively). The analysis identified 17 and 19 sites that were evolving faster in bacteria than eukaryotes, and vice versa, respectively. These 36 positions with the greatest rate differences were evaluated for their potential roles in the functional divergence of EFs by mapping them onto the known tertiary structures of bacterial EF-Tu.

The recognition of these 36 sites was based on a comparison of the observed rate differences for all 380 positions [with their mean and standard deviation (SD) of -0.03 and 1.52 replacements/site/unit evolutionary time, respectively] to their expected normal

distribution. These 36 sites were highlighted as those with rate differences of >2 SD from the mean. Although used in this way, these cutoffs were not viewed as rigorous thresholds of statistical significance, but were rather treated as conservative approximations with heuristic value. One obvious reason for this conservative interpretation was that the rate differences were not normally distributed, since they were based on a mixture of both stationary and non-stationary sites.

More importantly, we need to assume from the original estimates of replacement rates that the variances associated with the individual estimates get bigger for larger values of the rates themselves. Such trends in variance often occur because the error in the value being estimated is a percent of the value rather than an absolute value. Performing log-transformations on the original replacement rates can help to alleviate these biases.

When the underlying properties of a test statistic are poorly understood, computer simulations (parametric bootstrapping) offer one way to generate the null distributions for statistical testing (Huelsenbeck *et al.*, 1996). In this way, null distributions are derived from evolutionary simulations for the log-transformed rate differences between bacterial EF-Tu and eukaryotic EF-1 α . In these tests, the null model consists of a stationary gamma process, whereby the rate identities of sites remain constant over evolutionary time (Miyamoto and Fitch, 1995; Gu, 1999; Morozov *et al.*, 2000). As for the non-stationary covarion process, rate heterogeneity among sites is accommodated in this model by the gamma distribution (Yang, 1996). Thus, the only distinction between the stationary gamma and non-stationary covarion (gamma) approaches is that the latter allows for the evolutionary rates of sites to change across groups.

Log-Transformation Statistics

The same EFs, accepted phylogeny, ML approaches, and software used in the previous chapter were adopted in these simulations. The ML analyses relied on the Jones, Taylor, and Thornton matrix with rate variation among sites according to the gamma distribution (JTT-I') (Jones *et al.*, 1992; Yang, 1996). Evolutionary simulations involved two separate series of 100 simulations apiece, with each trial consisting of 30 simulated sequences of length 380 sites. In these simulations, α was set at 0.42 and 0.78, with the former representing the average of bacteria alone and eukaryotes alone. These simulations were to establish statistical thresholds for the observed rate differences.

Figures 3-1 and 3-2 show the relationships between non- and log-transformed rate differences for the simulated bacteria and eukaryotic data sets, respectively. To understand the importance of log-transforming the rate differences, consider the following example for two sites. The first site has a replacement rate of zero in bacteria and a rate of 2.5 in eukaryotes, whereas the second site has a rate of 1.5 in bacteria and 4.5 in eukaryotes. We can categorize these sites in evolutionary terms; site 1 is invariable in bacteria but moderately variable in eukaryotes, and site 2 is moderately variable in bacteria but highly variable in eukaryotes. Using the same statistical cutoffs as we did in the previous chapter (-3.0, 3.0), we would highlight site 2 only as having undergone functional divergence because it has a rate difference of 3.0 (site 1 has a difference of 2.5). However, it could be inferred that site 1 has functionally diverged because it has shifted between being variable and invariable. In fact, site 2 should probably not be highlighted because the variances associated with the rate estimates are larger, and may therefore overlap, than those associated with the rate of zero in site 1.

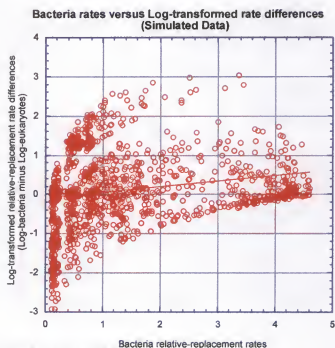
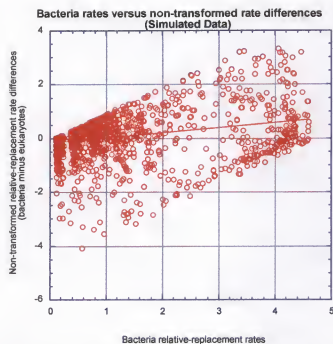


Figure 3-1. Comparison of simulated bacteria relative-replacement rates to non-transformed (top) and log-transformed (bottom) relative-replacement rate differences. Five random simulations (out of 100) were analyzed with alpha equal to 0.42 and branch lengths as determined by the EF ML calculations.

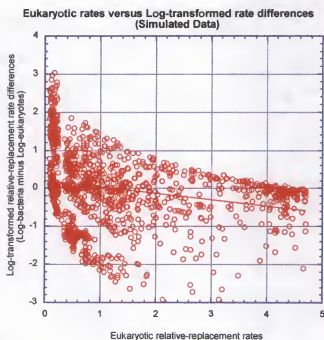
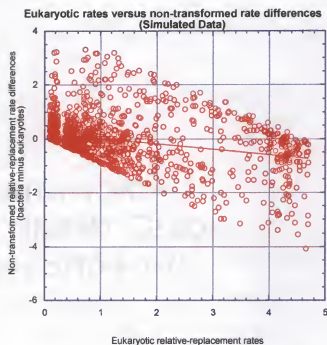


Figure 3-2. Comparison of simulated eukaryotic relative-replacement rates to non-transformed (top) and log-transformed (bottom) relative replacement rate differences. Five random simulations (out of 100) were analyzed with alpha equal to 0.42 and branch lengths as determined by the EF ML calculations.

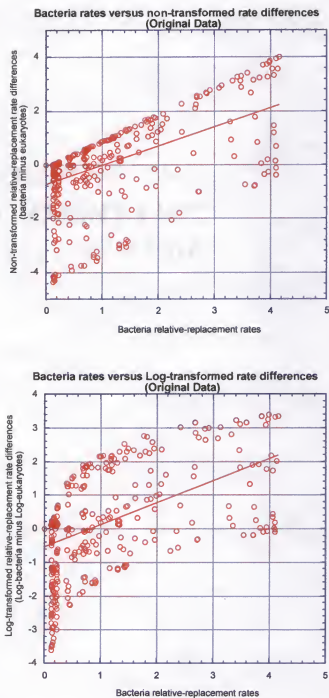


Figure 3-3. Comparison of original bacteria relative-replacement rates to non-transformed (top) and log-transformed (bottom) relative replacement rate differences.

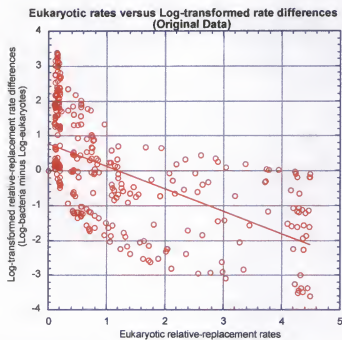
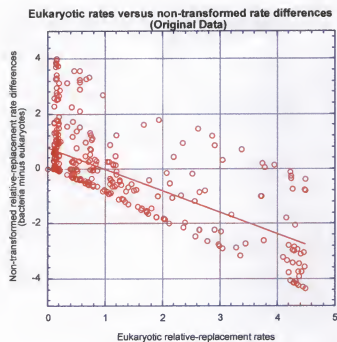


Figure 3-4. Comparison of original eukaryotic relative-replacement rates to non-transformed (top) and log-transformed (bottom) relative replacement rate differences.

Figure 3-1 shows how log-transforming the data can alleviate these confounding problems. We see from the top part of the figure that a cluster of sites at bacterial rates ~ 1.5 have rate differences around -3.0 . The corresponding sites in eukaryotes must then be around 4.5 . These sites can be interpreted as being moderately and rapidly evolving in bacteria and eukaryotes, respectively. Highlighting these sites as having undergone shifts in functional divergence, however, would be premature until we could analyze the variances associated with the rate estimates. We also notice in the top part of the figure that there are not many points on the graph that have a bacterial rate of ~ 0 rate with a corresponding rate difference near -3.0 . A shift in function should be most noticeable when a site has gone from invariable to variable, and vice versa. Based on the simulation conditions, most sites that shift between variable and invariable can be characterized as the variable site having a moderately evolving rate. Unfortunately, our previous method is not able to highlight these types of sites.

The problems associated with simply calculating rate differences can be overcome by log-transforming the data. The bottom part of Figure 3-1 highlights these adjustments. The sites that previously had rate differences near 3.0 , with the bacterial rate ~ 1.5 , have been 'pushed' towards the rate difference of zero. There is also an increase in the rate difference values for sites where the bacterial rate is ~ 0 . Thus, log-transformations eliminate some of the confounding effects of calculating non-transformed rate differences and they allow us to analyze the sites in a manner that is consistent with the notion of sites being 'independent and identically distributed.' These same adjustments take place for the eukaryotic sites of the simulated data, Figure 3-2, and for the original estimates of rates for bacteria and eukaryotic elongation factors (Figures 3-3 and 3-4, respectively).

The mean limits of the 100 frequency distributions with $\alpha = 0.42$ (-3.347, 3.269) were more conservative (i.e., encompassed a wider range) than those with $\alpha = 0.78$ (-2.535, 2.475). This relationship is predicted based on Figure 2-8 from the previous chapter. Higher alpha values represent a more uniform distribution of rates (standardized always to a mean of 1.0), whereas low values signify rate heterogeneity. Since the majority of sites in elongation factors have a conserved evolutionary pattern (slow rate), the limited numbers of rapidly evolving sites are *forced* to have an evolutionary rate similar to the conserved sites when alpha equals 0.78. Also, this relationship with α indicated that the rate differences of sites were not identically distributed, as the rapidly evolving positions were more heavily concentrated in the tails of their simulated frequency histograms than were the slowly evolving ones.

Before sites can be identified as evolving differently between bacterial and eukaryotic elongation factors, statistical thresholds first need to be calculated from the simulations with alpha equal to 0.42, Table 3-1. For the cutoffs, the 380 log-transformed rate differences were analyzed in all 100 trials individually. The means for the 100 trials were then used as the thresholds. For example, when $P=0.01$, 4 sites are analyzed (380×0.01)

Table 3-1. Statistical cutoffs based on the means for the 100 simulated trials (log-transformed bacterial rate minus log-transformed eukaryotic rate)

P value	Mean upper cutoff	Mean lower cutoff
0.01	2.489	-2.444
0.016	2.337	-2.283
0.021	2.221	-2.164
0.032	2.042	-1.985
0.042	1.930	-1.867

325	-3.604249739	x	x	x	x	x	p=0.01
189	-3.510145205	x	x	x	x	x	p=0.016
336	-3.481928716	x	x	x	x	x	p=0.021
69	-3.429568079	x	x	x	x	x	p=0.032
305	-3.375654627	x	x	x	x	x	p=0.042
103	-3.356042984	x	x	x	x	x	
138	-3.24914477	x	x	x	x	x	
288	-3.087438846	x	x	x	x	x	
326	-2.98309569	x	x	x	x	x	
106	-2.956153287	x	x	x	x	x	
323	-2.914004925	x	x	x	x	x	
144	-2.900147666	x	x	x	x	x	
160	-2.818398258	x	x	x	x	x	
350	-2.801938917	x	x	x	x	x	
167	-2.621038824	x	x	x	x	x	
96	-2.615912068	x	x	x	x	x	
98	-2.572553957	x	x	x	x	x	
136	-2.527052271	x	x	x	x	x	
337	-2.520524285	x	x	x	x	x	
298	-2.391562129	x	x	x	x	x	
354	-2.386555676	x	x	x	x	x	
282	-2.308567165	x	x	x	x	x	
277	-2.260940392		x	x	x	x	
133	-2.255588773		x	x	x	x	
151	-2.247385544		x	x	x	x	
99	-2.246862558		x	x	x	x	
150	-2.238540277		x	x	x	x	
68	-2.180268771		x	x	x	x	
190	-2.15350487			x	x	x	
315	-2.059238834			x	x	x	
134	-2.04731147			x	x	x	
72	-2.028716053			x	x	x	
94	-2.012159535			x	x	x	
4	-1.94866687				x	x	
287	-1.884034745				x	x	
335	-1.869924242				x	x	
322	-1.865380705				x	x	
205	-1.846523555						
347	-1.845447974						
361	-1.765902515						
7	-1.728983063						
324	-1.700369893						

Figure 3-5. Significant sites according to different statistical cutoffs [rates in the second column are calculated as $\text{Ln}(\text{Bac}) - \text{Ln}(\text{Euk})$] For purposes of space, sites 301 and 351 are not shown but highlighted when $P=0.042$. Sites are colored coded to their p values.

286	2.072472872				x	x	
241	2.082837131				x	x	
152	2.090006451				x	x	
270	2.149573365				x	x	
139	2.157219243				x	x	
171	2.1587972				x	x	
310	2.16521719				x	x	
331	2.170732962				x	x	
131	2.175099781				x	x	
295	2.180948488				x	x	
359	2.241139427		X		x	x	
28	2.285528691		X		x	x	
57	2.295150115			x	x	x	
356	2.298495108			x	x	x	
194	2.300917036			x	x	x	
226	2.337270651			x	x	x	
176	2.33972464	x	x		x	x	
228	2.356652314	x	x	x	x	x	
216	2.379546134	x	x	x	x	x	
50	2.394547429	x	x	x	x	x	
82	2.523621162	x	x	x	x	x	
64	2.571729044	x	x	x	x	x	
37	2.653522214	x	x	x	x	x	
39	2.692620773	x	x	x	x	x	
357	2.720929864	x	x	x	x	x	
355	2.747709126	x	x	x	x	x	
299	2.758539601	x	x	x	x	x	
263	2.804716313	x	x	x	x	x	
38	2.863880142	x	x	x	x	x	
31	2.965693764	x	x	x	x	x	
51	2.988906308	x	x	x	x	x	
271	3.015150359	x	x	x	x	x	
42	3.026037623	x	x	x	x	x	
163	3.08470931	x	x	x	x	x	
327	3.121760094	x	x	x	x	x	
40	3.148266741	x	x	x	x	x	
35	3.152121343	x	x	x	x	x	
123	3.234278973	x	x	x	x	x	
153	3.255075388	x	x	x	x	x	
36	3.328797352	x	x	x	x	x	
34	3.340431026	x	x	x	x	x	
311	3.353665685	x	x	x	x	x	
329	3.386021073	x	x	x	x	x	
Sites	Ln(bac)-Ln(euk)						
Mean	-0.019158113						
Stdev	2.537929149						
Max	3.386021073						
Min	-3.604249739						

in each trial. The site with the second largest (positive value) log-transformed rate difference and the site with the second smallest (negative value) difference are recorded for each trail. Therefore, a total of 100 positive and 100 negative rate differences are recorded. The means for the positive and negative numbers were calculated (Table 3-1) and now these means can be used as the cutoffs to analyze the elongation factor data set, Figure 3-5. A discussion of these results will be given in the next chapter.

Large rate differences are more likely for rapidly rather than slowly evolving sites. One primary implication of this interpretation is that the rate differences of the rapidly versus slowly evolving sites are not randomly distributed in their simulated frequency histograms. Instead, the rate differences for the former are more heavily represented in the tails of their distributions than are those of the latter. As a consequence, statistical thresholds (e.g., >2 SD in previous chapter) cannot be directly obtained from the tails of the non-transformed frequency distribution. Nevertheless, the log-transformation of this distribution resolve the biases associated with the correlation between large variances and large rates. Thus, by averaging across their 100 replicates, the mean minimums and maximums of these simulated frequency distributions establish valid cutoffs for the identification of covarion sites.

CHAPTER 4

CURRENT STATUS AND FUTURE PROSPECTS USING THE COVARION MODEL: BAYESIAN INFERENCE

Background

Genomic sequencing projects have provided scientists with an abundant amount of information to predict protein function (Bork and Koonin, 1998; Eisenberg *et al.*, 2000). Interpreting the information on a case-by-case basis and placing it within a biological context is a daunting task. Imagine biochemists and molecular biologists analyzing every gene from every sequencing project. Computational biology enables us to manage this information by making fundamental assumptions regarding the evolutionary process. The major annotation assumption is that orthologous sequences have the same function. Therefore, if we know the function of some protein in species A and subsequently sequence a gene from species B, and given an arbitrary statistical cutoff of sequence similarity between the two genes, then the protein in species B is said to have the same function as in species A. Recent studies, however, suggest that this underlying assumption may not hold true for many cases of orthologous sequences (see next chapter). At the core of these studies lies the notion that functional importance is highly correlated with conserved evolutionary sequence patterns. Although laboratory experiments are a necessary component to ultimately determine function, the goal of computational biology is to elucidate the function of genomic sequences as much as possible without having to perform these experiments. Functional genomics then seeks

to form a bridge between computational and experimental biology. Combining evolutionary and structural analyses with biochemical data enables the functional genomics field to make predictions regarding functional divergence. These predictions lead to direct testing of individual residues in the laboratory. Two approaches have recently been utilized that more accurately represent the underlying modes of divergent sequence evolution, and are therefore able to detect shifts along the functional continuum (nonsynonymous/synonymous ratios and covarian approaches). Due to their ability to detect changes in the patterns of sequence evolution, we predict that these approaches will become integral components of functional genomics studies in the future.

Evolutionary Tools for Functional Genomics

The most common way to detect functional divergence in genomic sequences is to calculate the nonsynonymous/synonymous rate ratio (K_a/K_s , d_N/d_S , or ω) (Yang and Bielawski, 2000). If a pair of sequences evolved under a neutral model of evolution then a comparison of the sequences will yield a nonsynonymous/synonymous ratio of 1 (e.g., pseudogenes). A pair of sequences under purifying selection will display a ratio less than 1, whereas sequences under diversifying selection display a ratio greater than 1. Until recently, methods calculated this ratio across all sites and therefore positive selection could only be detected if the average ratio across all sites was greater than 1. A serious limitation to these methods is evident. How can we detect positive selection within a background of otherwise purifying selection? Yang and colleagues have generated a method for detecting and identifying positive selection at individual sites. A Bayesian approach is used to calculate the probability that a site is under diversifying selection. Although this approach is powerful for identifying individual sites that are responsible for

functional divergence to subsequently be tested in the laboratory, the approach is not applicable to older evolutionary events as DNA becomes mutationally saturated at the third codon position more rapidly than the first and second positions. This sequence saturation will therefore lead to biases in the estimations of ω , compromising the ability to detect functional divergence.

The covarion (covariotide) model utilizes amino acid (DNA) data and can also be used to detect functional divergence among sequences. Originally proposed by Fitch and colleagues, the non-stationary covarion hypothesis allows the replacement rate at an individual position to vary in different branches of the evolutionary tree (Fitch and Markowitz, 1970; Miyamoto and Fitch, 1995). A site may be variable in one lineage but invariable in another lineage. This is in contrast to the stationary gamma model (Yang, 1996). Under this model the replacement rate can vary among positions within a sequence, like the covarion model, but the rate at any individual position must remain constant across all lineages. Previous studies have demonstrated that the covarion model can more accurately describe the mode of evolution for some proteins as compared to the gamma model (Lockhart *et al.*, 1998; Lopez *et al.*, 1999; Gaucher *et al.*, 2001). Although conceived nearly 30 years ago, a literature search will show that the majority of papers discussing the covarion hypothesis have been published within the past few years (Barbrook *et al.*, 1998; Finkelstein *et al.*, 1998; Tuffley and Steel, 1998; Gu, 1999; Lockhart *et al.*, 1999; Moreira *et al.*, 1999; Penny *et al.*, 1999; Philippe and Forterre, 1999; Collins *et al.*, 2000; Fisher *et al.*, 2000; Lockhart *et al.*, 2000; Naylor and Gerstein, 2000; Philippe and Germot, 2000; Philippe *et al.*, 2000; Steel *et al.*, 2000; Marin *et al.*, 2001). This is due in part to enhanced phylogenetic methods, computational speed, and a

greater need to understand molecular evolution in light of mass-sequencing projects. Fundamental to this understanding is the idea that shifts in sequence variability patterns can be indicative of shifts in protein function. Since the covarion model can detect these shifts in sequence variability, the model holds considerable promise for detecting functional divergence among various lineages. The remainder of this chapter will discuss recent advances using the covarion model and its utility for functional genomics.

Covarion Approaches: Methods of Overall Sequence Comparisons

Establishing a theoretical framework is an important first step when attempting to develop a method based on a proposed model. The covarion model is no exception. Tuffley and Steel have recently demonstrated this framework for reconstructing phylogenies (Tuffley and Steel, 1998). These authors show that the gamma and covarion models can be distinguished under specific conditions. Their results are based on the formulation of a distance measure that is tree additive under certain conditions using the covarion model [Kimura three-substitution (K3ST), Kimura two parameter (K2P) and Jukes-Cantor (JC)] but is not tree additive under the gamma model. Thus the covarion method can generate information from sequences that can ultimately enable a tree topology to be recovered quickly and uniquely.

Lockhart and colleagues have developed statistical tests that can determine if a set of sequences has evolved according to a gamma model or a covarion model (Lockhart *et al.*, 1998; Lockhart *et al.*, 2000). These statistics, contingency and inequality tests, were applied to data sets containing 16S rDNA and elongation factor (EF) sequences to elucidate the relationship of oxygenic photosynthetic lineages. For both data sets, the test statistics showed that the covarion model more accurately depicted the evolution of these

sequences as a whole compared to the gamma model. To determine the effects of covarion behavior on tree topology, the authors phylogenetically analyzed the 16S rDNA and EF sequences after sequentially removing the sites that displayed covarion behavior, or functional divergence (invariable in one lineage while variable in the other). This analysis revealed that the majority of bootstrap support for the original topology was in fact dependent on covarion sites. Therefore sites displaying non-stationary evolutionary patterns produced the most signal in support of the robust topology.

In a similar study concerned with phylogeny reconstructions, Philippe and colleagues analyzed the root of the Tree of Life in light of the covarion model using EFs (Lopez *et al.*, 1999). The root of the Tree of Life (or Universal Tree) is determined through the connectivity of two generated trees using anciently duplicated genes. Previous studies using EFs have suggested that the root lies on the branch separating bacteria from archaea and eukaryotes. In lieu of the fact that EFs are highly conserved but mutationally saturated at variable positions, Lopez *et al.* reanalyzed the placement of the root. These authors developed a method that utilizes parsimony to calculate the number of substitutions at each site and compares these estimates to a given threshold to increase the signal-to-noise ratio. A matrix was thus generated which contains information regarding the site-by-site relationships to the threshold. Incorporating this matrix into the tree building process resulted in a Universal Tree rooted on the branch separating eukaryotes from archaea and bacteria, albeit not robustly. These authors showed that the covarion model, when implemented in parsimony analyses, provides a better explanation of the evolutionary process for EFs than the gamma model. Both of the above studies suggest that shifts in evolutionary rates influence the resolved topologies. Since evolutionary

rates can be correlated to function, both studies suggest episodes of functional divergence have taken place within their evolutionary trees. Future work will need to determine when covarian behavior is phylogenetically informative and when it is misleading.

Covarian Approaches: Non-Baysian-based Methods for Identifying Sites

Although the above examples prove interesting for phylogeny reconstructions by analyzing sequences as a whole, studies that provide detailed information regarding individual sites within a covarian framework have proven more applicable to the field of functional genomics. Along these lines, Philippe and colleagues studied the relationships of eukaryotic lineages using EFs (Moreira *et al.*, 1999). By analyzing the site-by-site rate variation as estimated by parsimony, Moreira *et al.* showed that ciliates contain the largest number of variable positions for all eukaryotic lineages studied. In contrast, ciliates do not show such high variability in α -Tubulin or SSU RNA when compared to other lineages. The distribution of the sites displaying high variability in EFs was mapped along the primary sequence structure to identify whether any known protein motifs were affected by the increased variability. Motifs involved in translational activity, such as GDP-, GTP-, and aminoacyl-tRNA-binding, were equally conserved in ciliates as in other lineages. However, sites displaying high variability unique to ciliates were concentrated in regions that are putative actin-binding motifs. This is consistent with actin being a quantitatively minor protein and having an accelerated mutation rate in ciliates (c.f. Moreira *et al.*, 1999). Future work will determine if the high variability in these two molecules is due to a coevolutionary acceleration and/or a loss of their interactions.

The above study highlights the importance of incorporating the covarion model to analyze functional divergence. The authors were able to hypothesize that ciliates display uncharacteristically rapid sequence evolution in regions of EF that may be responsible for protein-protein interactions. Being able to assign individual sites as covarion-like afforded these authors the opportunity to attempt to correlate functional differences among eukaryotic lineages to specific sites in the EF sequence.

Incorporating maximum likelihood (ML) methods in phylogenetics can be advantageous when analyzing complex modes of evolution (Felsenstein, 1981). This is especially true when trying to identify specific sites of interest across diverse lineages for direct testing in the laboratory to determine functional divergence. As previously discussed, comparing the number of substitutions site-by-site between two lineages is necessary for determining covarion behavior within a functional genomics framework. Therefore, it is important to accurately estimate the number of expected substitutions for a site along a given topology. Parsimony may underestimate this parameter when the sequences have undergone parallel and/or back mutations. If the correct model is incorporated, ML is less susceptible to these confounding factors because the method incorporates branch lengths and an explicit substitution matrix into its calculations. In addition, incorporating the gamma distribution can provide a more accurate estimate of the site-by-site substitution rates.

The gamma model may be considered a subset of the covarion model. Both models allow rates to vary among sites. However, the gamma model requires that those rates remain the same (stationary) throughout evolutionary time, whereas the covarion model allows rates to fluctuate at individual positions between lineages throughout time (non-

stationary). Gu has mathematically demonstrated that if two lineages exhibit functional divergence, and if the functional divergence is ignored, then the estimation of the gamma distribution's shape parameter α , is biased (Gu, 1999).

Along these lines, we demonstrated that a covarion model explains the evolution of EFs between bacteria and eukaryotes more accurately than the gamma model (Gaucher *et al.*, 2001). Using ML methods, the estimated α value was 0.48 and 0.36 for the bacterial and eukaryotic groups, respectively. However, when the groups were combined the α value increased to 0.78. This shows that the evolutionary process for EFs is not stationary, as had it been stationary the α value would have been similar to the values obtained for the groups individually. Parametric bootstrapping simulations and sub-sampling experiments indicated that the α values were extremely robust to fluctuations. The non-stationary behavior of α for this data set indicated the importance of invoking the covarion model to explain the evolution of EFs.

As mentioned above, these types of covarion studies become powerful for functional genomics when individual sites are highlighted as displaying the most prominent covarion behavior and subsequently correlated to known functional/structural differences between the lineages under study. We used a histogram approach to identify covarion sites in EFs. The site-by-site replacement rates for bacteria were estimated by ML and compared to the site-by-site replacement rates for eukaryotes. A histogram was generated for the site-by-site rate differences between the two groups. The histogram was leptokurtotic, the mean and tails were over-represented while the shoulders were under represented, as compared to the expected distribution. Sites evolving under a stationary process had rate differences centered around the mean of zero. Sites evolving

according to an extreme non-stationary process had rate differences in the tails of the distribution. A total of 36 sites, out of 380, were highlighted as having a rate difference of >2 SD. The 36 sites were mapped onto the three-dimensional structures of EFs. This enabled us to generate testable hypotheses regarding known and putative functional/structural differences for EFs between bacteria and eukaryotes in the GDP-, GTP-, aminoacyl-tRNA-, and actin-binding regions (see Chapter 2). A functional genomics analysis was thus realized by correlating experimental and computational results.

Covarian Approaches: Bayesian-based Methods for Identifying Sites

Gu (99) has recently developed a method for detecting shifts in functional constraints between two or more gene clusters (lineages) as calculated from the expected number of replacements at individual sites. The model calculates the posterior probability for any site being in a state of functional constraint or functional divergence. The Bayesian approach has recently gained attention within the field of molecular evolution for its unique ability to calculate confidence intervals around estimated parameters, given a data set (Yang and Rannala, 1997; Larget and Simon, 1999; Lewis, 2001). These probabilistic statements cannot be generated by other likelihood methods that optimize the function by calculating the probability of observing the data set, given the parameters. It is assumed that when a site is invariable in one lineage but highly variable in a different lineage, the site has undergone functional divergence in one of the lineages compared to the ancestral state. This coefficient of functional divergence, θ , is calculated according to a maximum likelihood model that incorporates rate heterogeneity among sites and suggests whether a

significant number of sites have altered functional constraints between the analyzed gene clusters.

When the θ value indicates functional divergence, a value significantly greater than zero, a site-specific profile using a hidden Markov model (Bayesian) is generated to predict which sites are statistically most likely to be responsible for the divergence. This Bayesian-based approach calculates the posterior probability of a site being in a state of functional constraint (F_0) or functional divergence (F_1):

$$\Pr(B | A) = \frac{\Pr(B)\Pr(A | B)}{\Pr(A)}$$

Within the Bayesian framework, A represents the data while B represents a hypothesis (or state) and are integrated by the following:

$\Pr(B|A)$, posterior probability, is the probability of a site being in state F_0 or F_1 , given the data.

$\Pr(B)$, prior probability, is the unconditional probability of the hypothesis, represented by θ in this case.

$\Pr(A|B)$, likelihood, is the probability of the data, given the hypothesis.

$\Pr(A)$, unconditional probability of the data, used to standardize all possible probabilities so that the $\Pr(B|A)$ for all alternatives sum to 1. Thus, the probabilities of F_0 and F_1 sum to 1 for all individual sites.

Incorporating this method, Gu has predicted that functional divergence has taken place between mammalian transferrins and lactotransferrins, mammalian and non-mammalian transferrins, and between all combinations of N-, C-, and L-myc genes. In addition, the method has predicted which residues are most likely responsible for this functional divergence.

This approach contains formulations that can be advantageous over other methods. Calculating the posterior probability eliminates some of the problems associated with the existence of linear relationships between simply calculating site-by-site replacement differences between two lineages. For example, a site has 0 replacements in one lineage and 6 replacements in another lineage, while a different site has 9 replacements in the first lineage and 15 in the other. Both sites have a difference of 6 replacements between the two lineages, however only the former site should be highlighted as undergoing functional divergence. This is due to the former site shifting between variable and invariable, while the latter site is highly variable in both lineages. The variances associated with the estimated expected replacements in both lineages for the latter site are more likely to overlap and therefore negate the significance of 6 replacements. Accounting for this non-linear relationship of replacement differences greatly enhances the ability to detect covarion behavior as measured by functional divergence (see previous chapter).

We have applied Gu's method to analyze our EF data set (Figure 4-1). The θ value was 0.71 ± 0.03 and thus significantly greater than 0 and indicating functional divergence has taken place between bacterial and eukaryotic EFs. The site-specific profile highlighted 49 sites as having a posterior probability of 95% or greater and thus being in a state of functional divergence between the two lineages. Twenty-eight of these sites overlap with the sites highlighted by the original histogram method. A total of 24 sites were evolving more rapidly in eukaryotes than in bacteria. The majority of these sites are on the surface of the protein and are found in known protein- and nucleotide-binding domains. Alternatively, 25 sites were evolving more rapidly in bacteria than eukaryotes.

Eight of these sites lie in known binding domains, while 15 sites lie on the surface in areas with no known function. The majority of these sites do, however, lie in putative actin- and ribosome-binding domains. The increased evolutionary rate of these sites is consistent with actin being absent in bacteria and with the structural differences of the bacterial and eukaryotic ribosomes. Therefore, this approach of combining computational and experimental biology has predictive value regarding functional divergence.

In light of Gu's method being the most statistically sophisticated of all the current programs available for detecting functional divergence, we asked whether our log-transformed approach is comparable to his approach. We chose the threshold values that corresponded to $P=0.016$ from the simulations (see Table 3-1). Figure 4-2 compares the results using the original rate differences, Gu's methods, and log-transformed rate differences for the elongation factor data set. Both the log-transformed and Gu analyses highlight 49 sites as being the most statistically significant diverging positions between bacteria and eukaryotes using the arbitrary cutoffs of 98.4% and 95%, respectively. Surprisingly, thirty-seven of these 49 sites overlap. The ability of the log-transformed method to give results similar to Gu's method can be further analyzed. The importance of log-transformed rate differences is shown, for example, at positions 141 (row 55) and 288 (row 8). Position 288 is invariable in bacteria but moderately evolving in eukaryotes. It can therefore be inferred that this site has undergone functionally divergence because the selective constraints acting on this site have shifted. Due to the fact that the

(a)

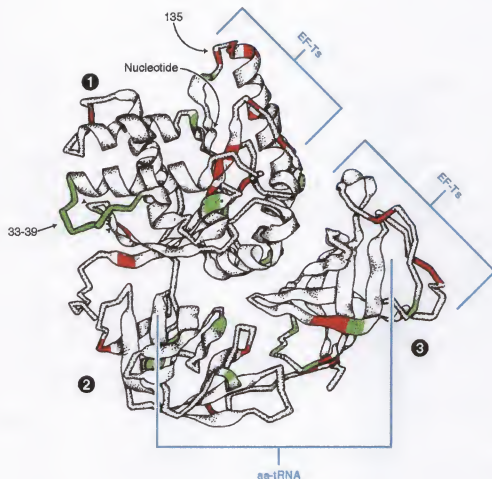


Figure 4-1. Functional divergence of bacterial and eukaryotic EFs. (a) Tertiary structure of the GTP-bound state for EF-Tu from *T. aquaticus* (Song *et al.*, 1999). Green and red highlight those 49 sites that have posterior probabilities greater than or equal to 95% according to Gu's method and are evolving faster in bacteria than in eukaryotes, and vice versa, respectively. (b) The known- and putative-behavioral roles of the sites highlighted in part A. The majority of these sites can be implicated for the functional and structural differences between bacterial and eukaryotic EFs in terms of translational and actin-binding activities.

(b)

Rapidly evolving in Eukaryotes only	Properties, function
4, 69, 160, 172, 288, 290, 350	Surface, no known function
106, 136, 144, 337	Surface, in proximity to EF-Ts and kirromycin binding
103, 133, 138, 336	Surface, all residues bind EF-Ts
189	On loop connecting domains 1 and 2
282, 325-326	Surface, 2-3 residues away from aa-tRNA binding
253, 277, 305	Surface, all residues bind aa-tRNA
96, 98	Interior, structural

Rapidly evolving in Bacteria only	Properties, function
33-39	Surface, possible localization sites or ribosome binding
131	310 Helix binds GTP/GDP, faces away from nucleotides
153, 163	Interior, no known function
51, 82, 203, 263, 271, 327, 329	Surface, 3-4 residues away from aa-tRNA binding sites
67, 123, 216, 286, 311, 335, 351, 357	Surface, possible localization sites

Figure 4-1 continued

original rate difference at this position is not very large, our original approach was not able to highlight this site. However, both the Gu and log-transformed approaches highlight this site as functional divergent between the two lineages. Alternatively, position 141 is moderately variable in both bacteria and eukaryotes. Since this site is at opposite ends of the 'moderately variable' spectrum, our original approach highlighted this site because of the large rate difference. Both the Gu and log-transformed methods agree in not highlighting this site. Thus these two approaches are comparable and outperform less-sophisticated methods for detecting functional divergence.

Figure 4-2. Comparison of three different methods for highlighting functionally divergent sites. The methods include our original approach for calculating rate differences, Gu's approach, and our log-transformed rate differences approach. The columns are represented by the following:

- A) The row number of this table.
- B) The site number corresponding to the EF multiple sequence alignment.
- C) The expected number of amino acid replacements for the bacterial lineage using Gu's program. This number is different, but similar, to the expected replacement rates that we calculate using our approach.
- D) The expected number of amino acid replacements for the eukaryotic lineage.
- E) The posterior probabilities according to Gu's method.
- F) Our original rate differences between the bacterial and eukaryotic lineages.
- G) Sites highlighted in our original analysis.
- H) Sites highlighted as having a posterior probability greater than 95%.
- I) Sites highlighted using the $P=0.016$ statistical thresholds generated from the **simulated** log-transformed rate differences.
- J) The log-transformed rate differences between bacteria and eukaryotic EFs.

1	325	0	15.54	0.99987	-4.362	X	X	X	-3.604249739
2	189	0	9.6406	0.99638	-4.154	X	X	X	-3.510145205
3	336	0	7.8765	0.9902	-4.224	X	X	X	-3.481928716
4	69	0	9.7388	0.99658	-4.151	X	X	X	-3.429568079
5	305	0	10.402	0.99765	-4.293	X	X	X	-3.375654627
6	103	0	11.181	0.99849	-4.179	X	X	X	-3.356042984
7	138	0	10.625	0.99793	-4.062	X	X	X	-3.24914477
8	288	0	5.6912	0.96689	-2.908		X	X	-3.087438846
9	326	0	10.214	0.99739	-4.106	X	X	X	-2.98309569
10	106	0	7.8706	0.99017	-2.442		X	X	-2.956153287
11	323	0	4.7125	0.94366	-2.632			X	-2.914004925
12	144	0	6.7017	0.98108	-2.817		X	X	-2.900147666
13	160	0	5.9262	0.97091	-3.15	X	X	X	-2.818398258
14	350	0	5.6516	0.96617	-1.981		X	X	-2.801938917
15	167	0	3.1335	0.87228	-2.754			X	-2.621038824
16	96	0	5.9933	0.97197	-1.623		X	X	-2.615912068
17	98	0	6.3321	0.97676	-1.585		X	X	-2.572553957
18	136	0	5.8896	0.97032	-1.739		X	X	-2.527052271
19	337	0	5.6912	0.96689	-1.498		X	X	-2.520524285
20	298	0	4.557	0.93877	-2.145			X	-2.391562129
21	354	0	2.1897	0.7998	-1.274			X	-2.386555676
22	282	0	7.2437	0.98602	-1.812		X	X	-2.308567165
23	277	1.0013	8.7365	0.95175	-3.918	X	X		-2.260940392
24	133	1.0269	10.861	0.98092	-3.758	X	X		-2.255588773
25	151	0	4.4269	0.93438	-1.828				-2.247385544
26	99	0	4.5665	0.93908	-1.108				-2.246862558
27	150	0	3.3016	0.88255	-1.81				-2.238540277
28	68	0	2.0154	0.78348	-1.193				-2.180268771
29	190	0	4.5792	0.9395	-1.523				-2.15350487
30	315	0	4.4396	0.93483	-1.368				-2.059238834
31	134	1.0372	6.8481	0.88965	-2.854				-2.04731147
32	72	0	3.3246	0.8839	-0.918				-2.028716053
33	94	0	3.1368	0.87249	-1.419				-2.012159535
34	4	1.0423	12.144	0.98935	-3.738	X	X		-1.94866687
35	287	0	4.3602	0.93202	-1.116				-1.884034745
36	335	0	2.1726	0.79824	-0.9				-1.869924242
37	322	2.2527	11.613	0.92667	-3.739	X			-1.865380705
38	205	0	2.2034	0.80104	-0.806				-1.846523555
39	347	1.0628	7.103	0.89688	-2.223				-1.845447974
40	361	0	2.1487	0.79605	-1.076				-1.765902515
41	7	0	2.1453	0.79573	-0.584				-1.728983063
42	324	1.0526	5.5936	0.82451	-1.956				-1.700369893
43	43	0	1.0372	0.67487	-0.568				-1.693319396

44	119	0	1.0443	0.67575	-0.563				-1.667274637
45	148	0	2.2068	0.80135	-0.634				-1.643076272
46	339	0	4.7157	0.94376	-0.829				-1.638025369
47	117	2.1501	6.9214	0.72651	-3.426	X			-1.624074623
48	253	2.1233	13.79	0.97324	-3.576	X	X		-1.606758741
49	135	2.1941	10.797	0.90777	-3.319	X			-1.595838753
50	1	2.1697	4.795	0.60253	-2.728				-1.56629842
51	290	2.1892	15.594	0.98636	-3.451	X	X		-1.559232166
52	70	0	2.0701	0.7887	-0.867				-1.558827912
53	16	0	2.1521	0.79636	-0.52				-1.556249601
54	307	1.0295	5.4044	0.81724	-1.419				-1.538417951
55	141	2.1965	8.5709	0.81314	-3.204	X			-1.519930263
56	367	0	3.4367	0.89027	-0.769				-1.506610968
57	140	1.0628	4.3253	0.74202	-2.214				-1.438258693
58	127	1.0269	5.6424	0.83148	-2.284				-1.424239293
59	303	0	2.2068	0.80135	-0.639				-1.36553237
60	79	0	1.0584	0.67751	-0.372				-1.356793697
61	266	0	1.0443	0.67575	-0.405				-1.298684719
62	320	0	1.0159	0.67221	-0.403				-1.295087593
63	207	0	1.0336	0.67443	-0.573				-1.275664733
64	250	0	1.0443	0.67575	-0.418				-1.240928511
65	62	0	1.0443	0.67575	-0.388				-1.213681618
66	102	4.6603	7.5542	0.43832	-3.092	X			-1.212568301
67	126	0	1.0301	0.67398	-0.309				-1.211577404
68	206	0	1.0336	0.67443	-0.306				-1.204735872
69	53	0	1.0336	0.67443	-0.508				-1.199854748
70	364	0	1.0336	0.67443	-0.349				-1.19272558
71	289	5.0009	8.0767	0.42095	-2.993				-1.191413407
72	341	0	1.0443	0.67575	-0.342				-1.183229337
73	11	0	1.0336	0.67443	-0.288				-1.178654996
74	149	3.2516	10.053	0.73768	-2.949				-1.177951184
75	21	0	2.1658	0.79762	-0.481				-1.171507003
76	158	3.4998	8.2171	0.59699	-2.601				-1.170926319
77	143	0	1.0372	0.67487	-0.285				-1.155487937
78	172	2.2209	12.864	0.95597	-3.025		X		-1.135878002
79	377	3.3282	12.613	0.85497	-2.934				-1.121388081
80	349	0	1.0372	0.67487	-0.435				-1.103231235
81	292	3.1705	10.322	0.76474	-2.84				-1.081084165
82	128	3.4094	8.5344	0.62717	-2.766				-1.073699132
83	122	0	1.0053	0.67088	-0.364				-1.036736885
84	105	0	1.0372	0.67487	-0.361				-1.031403539
85	61	0	0	0.5352	-0.202				-0.932945167
86	8	2.0964	3.1862	0.55186	-0.81				-0.910770939

87	227	1.0628	2.1658	0.60707	-0.636				-0.830870386
88	232	1.0244	3.2455	0.67493	-0.628				-0.778515558
89	156	1.0628	2.1042	0.60434	-0.795				-0.75728498
90	225	4.6295	5.5356	0.42183	-1.658				-0.732768977
91	332	2.1209	7.1938	0.7474	-0.99				-0.723236545
92	309	0	1.0584	0.67751	-0.232				-0.715419816
93	2	1.0295	2.1589	0.61017	-0.541				-0.696864657
94	265	0	1.0407	0.67531	-0.212				-0.683844788
95	281	1.0628	3.3181	0.67366	-0.589				-0.678209143
96	267	0	1.0407	0.67531	-0.21				-0.672385189
97	154	1.0269	4.6014	0.76693	-0.755				-0.671985763
98	147	3.4442	6.0941	0.51146	-1.046				-0.658692276
99	6	2.1404	4.63	0.59826	-0.677				-0.656170523
100	264	0	1.0443	0.67575	-0.2				-0.655406853
101	304	0	1.0089	0.67133	-0.202				-0.653561104
102	236	4.6317	14.039	0.75095	-2.059				-0.651479229
103	363	1.0628	2.1931	0.60833	-0.353				-0.623366183
104	318	0	1.0584	0.67751	-0.196				-0.614366303
105	328	4.6339	7.1763	0.43256	-1.161				-0.604898623
106	285	2.1013	3.3873	0.55596	-0.574				-0.591841568
107	195	5.8373	7.437	0.35846	-1.781				-0.560459314
108	352	1.0423	2.1213	0.60712	-0.473				-0.54873767
109	239	3.247	7.9233	0.61745	-0.978				-0.545840142
110	321	0	0	0.5352	-0.079				-0.499283863
111	201	2.1501	2.2034	0.54797	-0.477				-0.481609175
112	29	1.009	3.2554	0.67805	-0.415				-0.47865691
113	182	2.0964	4.4047	0.59312	-0.551				-0.472604411
114	23	0	0	0.5352	-0.069				-0.428454626
115	24	0	0	0.5352	-0.069				-0.428454626
116	166	0	0	0.5352	-0.069				-0.428454626
117	169	0	0	0.5352	-0.069				-0.428454626
118	46	0	0	0.5352	-0.065				-0.408045755
119	306	0	0	0.5352	-0.07				-0.407860319
120	101	2.1404	4.5856	0.59617	-0.36				-0.401795379
121	173	4.7416	7.5073	0.43041	-0.851				-0.388223302
122	280	2.1672	4.3602	0.58296	-0.396				-0.379858036
123	60	0	0	0.5352	-0.055				-0.333384226
124	183	4.7504	5.9079	0.41446	-0.774				-0.333255154
125	257	0	0	0.5352	-0.051				-0.333144447
126	175	5.9868	5.6973	0.38069	-1.005				-0.315430043
127	348	5.9951	8.992	0.36011	-0.999				-0.312231519
128	273	4.4515	5.8194	0.42987	-0.515				-0.277236892
129	297	4.7592	5.8743	0.41416	-0.401				-0.258732437

130	243	0	0	0.5352	-0.031				-0.226416877
131	230	2.1086	3.3246	0.55389	-0.215				-0.199456593
132	379	9.1414	14.813	0.23531	-0.774				-0.189390621
133	302	0	0	0.5352	-0.034				-0.188400603
134	168	7.7178	11.503	0.2783	-0.728				-0.186970435
135	346	9.4454	13.18	0.2017	-0.75				-0.183800151
136	242	1.0397	1.0018	0.58467	-0.087				-0.183478295
137	26	0	0	0.5352	-0.013				-0.101253734
138	256	11.3	20.608	0.20043	-0.387				-0.090134288
139	113	2.1892	2.2273	0.54673	-0.065				-0.087480009
140	251	6.2752	11.259	0.38813	-0.347				-0.085446314
141	231	3.3885	3.4894	0.49315	-0.095				-0.083053387
142	188	3.3398	4.6871	0.48926	-0.136				-0.067890336
143	254	4.7108	5.7828	0.41679	-0.16				-0.065921721
144	66	9.2806	10.7	0.21061	-0.269				-0.065657345
145	240	2.0989	2.0837	0.55147	-0.041				-0.056999444
146	124	2.1453	2.135	0.54955	-0.047				-0.056268582
147	342	2.1941	2.0188	0.55225	-0.033				-0.042178776
148	224	11.481	12.247	0.14847	-0.125				-0.030061179
149	196	1.0423	1.0159	0.58444	-0.014				-0.028690492
150	362	1.0423	1.0336	0.58418	-0.009				-0.021609484
151	110	1.0244	1.0584	0.58403	-0.01				-0.020161973
152	65	1.0397	1.0336	0.5842	-0.005				-0.012706651
153	375	1.0526	1.0443	0.58398	-0.002				-0.004773279
154	107	0	0	0.5352	0.002				0.010050336
155	370	1.0244	1.0584	0.58403	0.007				0.016260521
156	192	7.9346	6.9009	0.31737	0.064				0.016824792
157	174	11.526	8.9415	0.23028	0.145				0.036113998
158	10	0	0	0.5352	0.005				0.04184711
159	217	0	0	0.5352	0.005				0.04184711
160	187	10.651	9.2363	0.2187	0.218				0.056756511
161	162	2.0281	3.2521	0.55845	0.048				0.058339749
162	137	5.9868	7.4399	0.35136	0.311				0.097523563
163	291	0	0	0.5352	0.021				0.099372474
164	12	0	0	0.5352	0.021				0.100804699
165	17	0	0	0.5352	0.021				0.100804699
166	77	0	0	0.5352	0.021				0.100804699
167	88	0	0	0.5352	0.021				0.100804699
168	104	0	0	0.5352	0.021				0.100804699
169	211	0	0	0.5352	0.021				0.100804699
170	213	0	0	0.5352	0.021				0.100804699
171	223	0	0	0.5352	0.021				0.100804699
172	258	0	0	0.5352	0.021				0.100804699

173	262	0	0	0.5352	0.021				0.100804699
174	276	0	0	0.5352	0.021				0.100804699
175	340	0	0	0.5352	0.021				0.100804699
176	373	0	0	0.5352	0.021				0.100804699
177	3	0	0	0.5352	0.017				0.105360516
178	18	0	0	0.5352	0.017				0.105360516
179	76	0	0	0.5352	0.02				0.105360516
180	83	0	0	0.5352	0.017				0.105360516
181	130	0	0	0.5352	0.017				0.105360516
182	191	0	0	0.5352	0.02				0.105360516
183	114	0	0	0.5352	0.022				0.107420249
184	115	0	0	0.5352	0.022				0.107420249
185	181	0	0	0.5352	0.022				0.107420249
186	252	0	0	0.5352	0.022				0.107420249
187	49	0	0	0.5352	0.017				0.109433841
188	111	0	0	0.5352	0.017				0.109433841
189	146	0	0	0.5352	0.017				0.109433841
190	221	0	0	0.5352	0.017				0.109433841
191	246	0	0	0.5352	0.017				0.109433841
192	85	0	0	0.5352	0.014				0.110348057
193	247	0	0	0.5352	0.014				0.110348057
194	52	0	0	0.5352	0.016				0.111225635
195	193	0	0	0.5352	0.016				0.111225635
196	275	0	0	0.5352	0.016				0.111225635
197	15	0	0	0.5352	0.016				0.112005058
198	44	0	0	0.5352	0.016				0.112005058
199	74	0	0	0.5352	0.016				0.112005058
200	80	0	0	0.5352	0.016				0.112005058
201	93	0	0	0.5352	0.016				0.112005058
202	132	0	0	0.5352	0.016				0.112005058
203	186	0	0	0.5352	0.016				0.112005058
204	165	0	0	0.5352	0.012				0.112267302
205	91	0	0	0.5352	0.016				0.112795494
206	108	0	0	0.5352	0.016				0.112795494
207	121	0	0	0.5352	0.014				0.113023388
208	215	0	0	0.5352	0.014				0.113023388
209	278	0	0	0.5352	0.014				0.113023388
210	5	0	0	0.5352	0.015				0.114192368
211	13	0	0	0.5352	0.015				0.114192368
212	78	0	0	0.5352	0.015				0.114192368
213	112	0	0	0.5352	0.015				0.114192368
214	95	0	0	0.5352	0.014				0.114880276
215	54	0	0	0.5352	0.014				0.115831816

216	56	0	0	0.5352	0.014				0.115831816
217	209	0	0	0.5352	0.014				0.115831816
218	84	0	0	0.5352	0.014				0.117783036
219	129	0	0	0.5352	0.014				0.117783036
220	260	0	0	0.5352	0.014				0.117783036
221	20	0	0	0.5352	0.014				0.121889818
222	22	0	0	0.5352	0.014				0.121889818
223	55	0	0	0.5352	0.014				0.121889818
224	87	0	0	0.5352	0.014				0.121889818
225	109	0	0	0.5352	0.014				0.121889818
226	214	0	0	0.5352	0.014				0.121889818
227	369	0	0	0.5352	0.014				0.121889818
228	157	2.1648	3.4894	0.5534	0.157				0.12491386
229	368	0	0	0.5352	0.018				0.126040721
230	100	4.6427	3.407	0.49299	0.161				0.131564394
231	63	0	0	0.5352	0.03				0.139112802
232	81	0	0	0.5352	0.03				0.139112802
233	353	4.8317	3.3115	0.50322	0.189				0.152653674
234	92	0	0	0.5352	0.019				0.152907672
235	89	0	0	0.5352	0.021				0.177681177
236	272	5.9805	5.8011	0.37752	0.434				0.180496739
237	378	3.4511	2.2034	0.55779	0.172				0.187709178
238	316	0	0	0.5352	0.026				0.190353728
239	179	8.6049	5.9628	0.37393	0.716				0.19218691
240	365	0	0	0.5352	0.029				0.194591649
241	235	0	0	0.5352	0.039				0.196114879
242	204	0	0	0.5352	0.03				0.202026628
243	274	3.4279	3.3543	0.49586	0.264				0.225246602
244	59	4.902	3.3378	0.50267	0.27				0.226146557
245	90	0	0	0.5352	0.028				0.244691888
246	177	7.4678	4.4174	0.4785	0.86				0.259629374
247	47	0	0	0.5352	0.047				0.267879445
248	212	0	0	0.5352	0.038				0.287682072
249	161	4.6603	1.0372	0.78594	0.294				0.298569336
250	210	3.3978	1.0443	0.69555	0.26				0.313930299
251	300	0	0	0.5352	0.046				0.315516871
252	238	7.5889	5.6027	0.38396	0.94				0.336593862
253	229	4.5944	3.3774	0.4943	0.474				0.337682158
254	170	1.0628	1.0372	0.58404	0.204				0.346276237
255	344	14.407	5.145	0.74465	1.192				0.348183367
256	48	0	0	0.5352	0.049				0.354934299
257	268	3.1867	2.1897	0.55212	0.347				0.355076616
258	125	3.4395	2.2034	0.55746	0.314				0.357358705

259	142	0	0	0.5352	0.069				0.384845821
260	249	2.1404	1.0584	0.61332	0.298				0.436830767
261	255	1.0423	1.0089	0.58455	0.244				0.438764522
262	237	10.969	6.0025	0.43941	1.45				0.441832752
263	164	1.0269	1.0372	0.58424	0.253				0.471490618
264	261	2.1721	1.0584	0.61496	0.296				0.526625209
265	116	2.177	1.0443	0.61675	0.298				0.530352046
266	185	3.2354	1.0089	0.68974	0.476				0.544301553
267	284	4.7811	2.1008	0.62773	0.541				0.54703944
268	180	4.4669	3.295	0.49803	0.868				0.585745848
269	366	0	0	0.5352	0.099				0.601579987
270	313	0	0	0.5352	0.107				0.622136145
271	9	0	0	0.5352	0.102				0.626897795
272	371	0	0	0.5352	0.104				0.644139601
273	245	4.5548	2.1692	0.60676	0.691				0.649838653
274	159	14.349	5.9262	0.62497	1.794				0.65802806
275	317	0	0	0.5352	0.108				0.666478933
276	380	7.3271	3.3873	0.59359	1.61				0.675313853
277	293	4.601	2.1453	0.61216	0.726				0.682306966
278	279	4.5504	3.2488	0.50249	1.227				0.705600642
279	296	0	0	0.5352	0.117				0.706219262
280	233	1.0295	0	0.67999	0.217				0.732176524
281	343	6.1631	2.1008	0.71164	1.046				0.857253695
282	319	3.4998	1.0301	0.7052	1.124				0.898517133
283	334	4.6449	1.0372	0.78491	0.957				0.904236808
284	330	2.1233	1.0584	0.61245	0.769				0.930166078
285	374	4.6691	2.2034	0.60838	1.238				0.977309008
286	155	4.6163	1.0443	0.78176	1.306				1.0239437
287	338	1.0295	0	0.67999	0.275				1.028685274
288	333	1.0295	0	0.67999	0.455				1.193310099
289	218	1.0295	0	0.67999	0.493				1.249872793
290	372	1.0423	0	0.68165	0.295				1.258849415
291	358	1.0372	0	0.68098	0.491				1.261560679
292	97	1.0526	0	0.68297	0.494				1.265930679
293	86	1.0526	0	0.68297	0.294				1.275068726
294	219	1.0295	0	0.67999	0.354				1.281750505
295	360	1.0244	0	0.67932	0.355				1.283789242
296	199	1.0295	0	0.67999	0.509				1.287498733
297	197	1.0244	0	0.67932	0.402				1.288530192
298	220	6.1071	1.0089	0.87508	1.182				1.318045261
299	294	4.5548	1.0159	0.78248	1.361				1.321621151
300	376	7.4932	1.0584	0.925	1.525				1.325131778
301	27	8.8887	2.2273	0.84951	2.692				1.334563286

302	71	1.009	0	0.67733	0.304				1.338892122
303	19	1.0628	0	0.68429	0.288				1.348458827
304	314	1.0628	0	0.68429	0.379				1.506132896
305	118	1.0423	0	0.68165	0.385				1.518377947
306	222	1.0526	0	0.68297	0.406				1.560092038
307	45	5.827	1.0443	0.8557	2.106				1.568916715
308	283	1.0295	0	0.67999	0.596				1.588301062
309	200	1.0423	0	0.68165	0.539				1.602057805
310	345	6.0117	1.0124	0.86988	1.721				1.609902921
311	312	2.1111	0	0.80177	0.995				1.716356846
312	33	9.3447	1.0372	0.96867	3.338	X	X		1.725775306
313	248	1.009	0	0.67733	0.574				1.727937537
314	41	4.6031	0	0.94688	1.562				1.738831628
315	203	4.8514	0	0.95384	0.935		X		1.761229405
316	120	2.1453	0	0.80499	0.959				1.765318696
317	269	7.376	2.1589	0.77748	3.283	X			1.78117736
318	202	2.1233	0	0.80293	0.899				1.790833114
319	58	2.1892	0	0.80906	0.572				1.794679179
320	308	2.0769	0	0.79852	0.904				1.795456331
321	178	5.912	1.0443	0.8602	2.908				1.818918388
322	259	2.1355	0	0.80407	0.698				1.819758864
323	208	3.1867	0	0.88469	0.795				1.823916581
324	14	2.1843	0	0.80861	0.617				1.836335094
325	234	8.042	1.0053	0.94525	2.239				1.83946082
326	73	2.2161	0	0.81152	0.612				1.851351566
327	25	2.1428	0	0.80476	0.627				1.871802177
328	30	2.011	0	0.79214	0.63				1.875842586
329	198	2.1233	0	0.80293	1.009				1.887911046
330	32	7.4541	1.0301	0.92649	3.146	X			1.898923412
331	244	1.0628	0	0.68429	0.578				1.905500611
332	75	1.0269	0	0.67966	0.66				1.907930901
333	67	9.4307	1.0584	0.96891	3.259	X	X		1.930551263
334	184	2.1453	0	0.80499	0.696				1.953335926
335	145	3.3004	0	0.89141	0.946				1.971718033
336	301	3.3143	0	0.89221	1.116				1.974081026
337	351	14.175	1.0372	0.99726	3.566	X	X		2.008007693
338	286	4.9196	0	0.95559	1.25		X		2.072472872
339	241	3.1264	0	0.88097	1.033				2.082837131
340	152	4.6119	0	0.94714	1.084				2.090006451
341	270	2.011	0	0.79214	0.887				2.149573365
342	139	2.1599	0	0.80635	1.04				2.157219243
343	171	2.0793	0	0.79876	0.858				2.1587972
344	310	2.2527	0	0.81483	1.497				2.16521719

345	331	2.1428	0	0.80476	1.056				2.170732962
346	131	9.0169	1.0336	0.96361	3.566	X	X		2.175099781
347	295	3.4279	0	0.89854	0.919				2.180948488
348	359	3.4604	0	0.90029	1.664				2.241139427
349	28	3.414	0	0.89778	1.201				2.285528691
350	57	4.6779	0	0.94907	1.205				2.295150115
351	356	3.4001	0	0.89702	1.317				2.298495108
352	194	4.5438	0	0.94507	1.617				2.300917036
353	226	3.2656	0	0.88939	1.431				2.337270651
354	176	3.4836	0	0.90152	3.123	X		X	2.33972464
355	228	3.3096	0	0.89194	1.892			X	2.356652314
356	216	5.829	0	0.97366	1.764		X	X	2.379546134
357	50	4.6339	0	0.94779	1.355			X	2.394547429
358	82	4.746	0	0.95099	1.308		X	X	2.523621162
359	64	4.6207	0	0.9474	1.777			X	2.571729044
360	37	5.8311	0	0.9737	2.654		X	X	2.653522214
361	39	8.1748	0	0.99335	2.878		X	X	2.692620773
362	357	6.5076	0	0.98226	2.555		X	X	2.720929864
363	355	6.4329	0	0.98147	1.709		X	X	2.747709126
364	299	4.6163	0	0.94727	1.655			X	2.758539601
365	263	6.4205	0	0.98133	3.12	X	X	X	2.804716313
366	38	5.7502	0	0.97243	2.529		X	X	2.863880142
367	31	4.2889	0	0.93665	2.706			X	2.965693764
368	51	5.8788	0	0.97441	2.773		X	X	2.988906308
369	271	5.8332	0	0.97373	2.967		X	X	3.015150359
370	42	4.5614	0	0.94561	2.295			X	3.026037623
371	163	10.798	0	0.9986	3.755	X	X	X	3.08470931
372	327	10.299	0	0.99811	3.318	X	X	X	3.121760094
373	40	4.5328	0	0.94473	2.564			X	3.148266741
374	35	9.1945	0	0.99636	3.425	X	X	X	3.152121343
375	123	5.9951	0	0.97608	3.268	X	X	X	3.234278973
376	153	9.0444	0	0.99603	3.813	X	X	X	3.255075388
377	36	11.72	0	0.99919	3.955	X	X	X	3.328797352
378	34	12.589	0	0.99952	4.003	X	X	X	3.340431026
379	311	11.816	0	0.99923	3.727	X	X	X	3.353665685
380	329	9.7383	0	0.99737	3.854	X	X	X	3.386021073
row #	Site	Bacsubs	Euksubs	Posteriors	Rate diff	ORI	Gu	Log-trans	Ln(bac)-Ln(euk)
	Mean				-0.03175	-3.07	95%	-2.2829	0.108655438
	Stddev				1.51868	3.01		2.3373	1.47113912
	Max				4.003				3.386021073
	Min				-4.362				-3.604249739
	N				380	36	49	49	380

In all, we believe the above analyses demonstrate the well-rounded initiation of a functional genomics study. Evolutionarily divergent sequence patterns were highlighted and mapped onto the three-dimensional structure of EF and subsequently correlated with known biochemical differences between the two lineages under study. The successful completion of a functional genomics study will now require site-mutagenesis studies to determine if the sites displaying divergent evolutionary patterns are responsible for some or all of the biochemical/behavior differences of EFs.

Predictive Power of Covarion Analyses

Based on Figure 4-1, we would predict that the binding mechanisms of nucleotide exchange factors to their respective EFs are not equivalent in eukaryotes and bacteria. This hypothesis is formulated based on the fact that eukaryotes display rapid replacement rates at many positions that are correspondingly conserved in bacteria, and these positions have been shown to bind the nucleotide exchange factor in the latter lineage. Recent experimental evidence supports our hypothesis. The crystal structure of eukaryotic yeast EF-1 α bound to its nucleotide exchange factor has been determined at the 1.67 Å level (Andersen *et al.*, 2000).

The overall structure of eukaryotic EF-1 α is very similar to the bacterial EF-Tu, as predicted based on the high degree of sequence conservation between the two lineages. However, unforeseen to experimental biologists but predicted by our comparative evolutionary approach, the eukaryotic nucleotide exchange factor does not bind the same EF regions as demonstrated for EF-Tu:EF-Ts. Figure 4-3 shows that EF-1 α binds its nucleotide exchange factor predominately at the interface of domains 1 and 2.

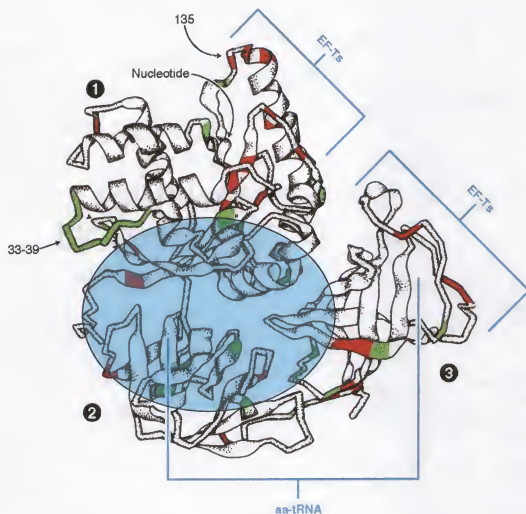


Figure 4-3. Crystal structure of EF-Tu with substrate binding domains labeled. The binding domain of nucleotide exchange factor and EF-1 α is highlighted in transparent blue. Note, the binding domains for the nucleotide exchange factor are significantly different between the bacterial EF-Tu and eukaryotic EF-1 α .

Surprisingly, this is the same region that binds tRNA. In fact, many of the residues in EF-1 α that bind tRNA also bind the nucleotide exchange factor. Further investigation shows that three EF positions are slowly evolving in eukaryotes but rapidly evolving in bacteria (covarian behavior) and are in regions that form hydrogen bonds between the exchange factor and EF-1 α domain 2. As EF-1 α has been shown to interact with tRNA synthetases via the exchange factor, along with the fact that eukaryotic tRNA is not free in solution and EF-1 α can bind actin, these results support the idea that eukaryotic EFs are involved in 'tRNA-channeling' from the nucleus to the ribosome. This functional annotation cannot be applied to EF-Tu since bacteria lack actin and a nucleus.

In addition to the above example, the recent crystal structure data suggest that EF-1 α does not undergo major conformational shifts between the GTP- and GDP-bound states, as seen in Figure 2-6. This was predicted based on experimental data (Negrutskii and El'skaya, 1998) and by our analysis. Residues 33-39 are rapidly evolving in bacteria and are thought to act as a hinge thereby allowing domain 1 to re-orient itself relative to domains 2 and 3 between the GTP- and GDP-bound states in bacteria. In fact, the crystal structure of this stretch of residues has never been elucidated due to high resonance and therefore does not display a characteristic secondary structure such as an alpha helix and beta strand in EF-Tu. This stretch of residues is conserved in EF-1 α and is an alpha helix, as predicted by us (c.f. Gaucher *et al.*, 2001). Using EF-Tu:GTP as a reference point, domain 1 is rotated by 25° relative to domains 2 and 3 in EF-1 α bound to the nucleotide exchange factor, and 83° in EF-Tu:GDP and EF-Tu:EF-Ts (Kawashima *et al.*, 1996; Andersen *et al.*, 2000). This large degree of rotation prevents EF-Tu from binding tRNA in the GDP bound state. We can assume the structure of EF-1 α :GTP is similar to

the structure of EF-1 α bound to the exchange factor, and thus similar to EF-1 α :GDP, because EF-1 α can bind to tRNA in both the GTP- and GDP-bound states (Negrutskii and El'skaya, 1998). It would therefore appear our prediction is correct concerning the lack of major conformational shifts in eukaryotic EF-1 α . We would suggest that it might be possible to synthesize an organic molecule that can bind in the large pocket unique to the EF-Tu:GDP state without binding to EF-1 α . This may prevent EF-Tu from recharging and thus act as an antibiotic.

These recent data strongly support some of the hypotheses previously generated (Gaucher *et al.*, 2001). Thus, incorporation of the covarion model into comparative evolutionary studies holds considerable promise for detecting functional divergence among proteins at the level of sequence analysis.

Covarion Approaches, Evolutionary Tools, and Functional Genomics

Functional divergence can be detected by changes in evolutionary rates (Messier and Stewart, 1997; Golding and Dean, 1998; Chang and Donoghue, 2000). For the covarion model, functional divergence is best detected when a site is conserved in one lineage but highly variable in another lineage. This means that the site has gained a functional role in one lineage, or lost its role in the other, compared to the ancestral state. However, the covarion model cannot detect all episodes of functional divergence. When a site has a specific function and is highly conserved in the ancestral state and one descendent lineage retains the functional state while another descendent lineage loses the functional state only to gain a new function later, the covarion approach cannot identify this change in functional behavior because the site is now conserved in both lineages. This behavior is exemplified by the EF results. The large helix in the upper right hand corner of domain

1 in EF-Tu binds the nucleotide exchange factor EF-Ts in bacteria, whereas EF-1 α does not bind its exchange factor here (Figure 4-3). Based on sequence similarity, it is hypothesized that this region in EF-1 α binds actin (Yang *et al.*, 1990). Thus, these positions are conserved in both lineages (invariant but different) yet they perform different functions. Although limited in this way, the covarion model holds considerable promise for future studies. In addition, the non-stationary behavior of parameters in phylogenetic analyses is beginning to gain considerable attention. Whether analyzing codon-usage (Sharp and Matassi, 1994), evolutionary distance (Gu and Li, 1998), rate variation (Grishin *et al.*, 2000; Morozov *et al.*, 2000), or base composition (Mooers and Holmes, 2000), non-stationary models are proving to be more accurate at representing the dynamic evolutionary process compared to stationary versions.

The covarion approach now offers an alternative to the nonsynonymous/synonymous approach for detecting functional divergence in genomic sequences. However, we do not view these approaches as being mutually exclusive. Functional genomics analyses are enhanced when these two approaches are used in concert. Both methods can be reliable for determining recent evolutionary events resulting in functional divergence (although the covarion model is predicted to work better than ω for older evolutionary events). As demonstrated for EFs and by Yang, both methods can detect functional divergence within a background of conserved sequence evolution. In this review, we have shown that covarion behavior can be detected using multiple approaches (statistical tests, parsimony, and maximum likelihood) and that individual sites can be highlighted using multiple approaches (variable/invariable, histogram, Markov/Bayesian). Although the covarion hypothesis is difficult to model, its incorporation into phylogenetic analyses has furthered

our understanding of molecular evolution. Overall, we advocate the use of evolutionary approaches for determining protein function. Simple statistical tests (e.g., BLAST, Lipman and Pearson, 1985) fail to identify important historical events that led to functional divergence. Therefore, functional genomics studies are at their most powerful when more realistic evolutionary models are incorporated into the computational aspects of the study. We predict that functional genomics will utilize sophisticated, yet easily manageable, evolutionary tools to infer protein function in the future.

CHAPTER 5 EVOLUTION, LANGUAGE AND ANALOGY IN FUNCTIONAL GENOMICS

Background

Almost a century ago, Wittgenstein pointed out that theory in science is intricately connected to language. This connection is not a frequent topic in the genomics literature. But a case can be made that functional genomics is today hindered by the paradoxes that Wittgenstein identified. If this is true, until these paradoxes are recognized and addressed, functional genomics will continue to be limited in its ability to extrapolate information from genomics sequences.

Those who ask "What is the function of my protein?" expect a linguistic answer (Wittgenstein, 1993), a sentence or two written in the language of the biologist. The answer might take, as an example, the form: "Your protein is a leptin, which regulates the feeding behavior of mice. When the gene is mutated or deleted, the mouse becomes obese" (Zhang *et al.*, 1994).

How does one get such a linguistic construct from a genomic sequence, which is no more (and no less) than a chemical formula for an organic molecule? This question, central to contemporary functional genomics, is not easy to answer. The simpler task of predicting how the behavior (not function) of an organic molecule is determined by its structure remains one of the great unsolved problems in chemistry. In principle, we should be able to solve this problem. The "First Law of Chemistry" states that the behavior of *all* matter is determined by the behavior of its constituent molecules, even behavior that a biologist might observe and call a phenotype. However, this has not been done convincingly for any but the simplest of molecules, and we are far from doing it for the general molecule, let alone a protein.

And even if we could do so, behavior would not necessarily lead to a statement about function. For example, it might become predictable that the benzodiazapene receptor binds tightly to valium. But the implied statement, "the purpose of this receptor is to bind to valium", is transparently mis-derived because valium is synthetic. To go from molecular behavior to organismic fitness, which is the Darwinian definition of function, information is required about the entire organism and the entire ecosystem.

"Functional Equivalency"

To obtain functional annotation, contemporary bioinformatics generally attempts to bridge chemical sequence to biological fitness using a doctrine of "functional equivalency" (for example, see Eisenberg *et al.*, 2000). This doctrine seeks to write a linguistic construct for a new protein sequence by expropriating the linguistic construct from another sequence having a similar chemical structure, under the assumption that the two proteins with similar chemical structures have equivalent functions. A protein with unknown function is found in one genome. It is inferred, from its sequence similarity, to be homologous to a different protein found in a different organism. Homologous proteins are then assumed to have equivalent functions. The functional language assigned to the protein with the known function is then transferred to the new protein.

Long before the genomics revolution began, many cases were known where this doctrine failed (Benner and Ellington, 1988). Figure 5-1 illustrates just one example. Here, four proteins from microbial metabolism, adenylosuccinate lyase, argininosuccinate lyase, aspartase, and fumarase clearly group into homologous pairs based on sequence similarity, and are part of an evolutionary superfamily that includes all four proteins (Aimi *et al.*, 1990). One protein is involved in nucleic acid biosynthesis, another is involved in amino acid biosynthesis, another is involved in amino acid degradation, and the last is involved in central metabolism, however. The biologist certainly does not regard the function of these proteins as equivalent.

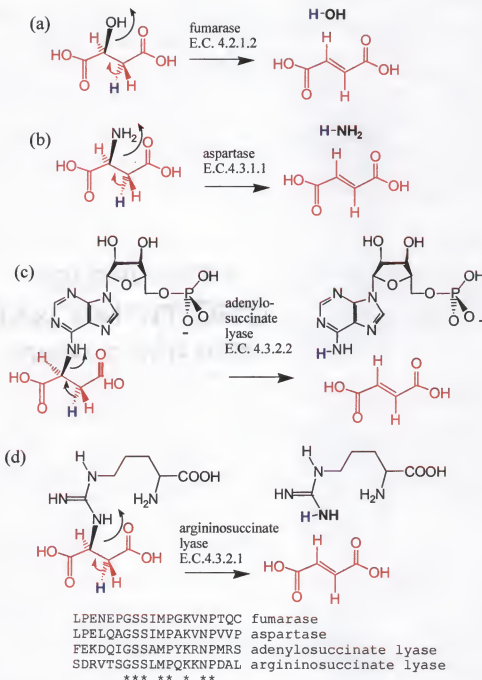


Figure 5-1. Using analogy to determine function. Homologous enzymes catalyze four reactions: (a) in central metabolism (the citric acid cycle) (b) in amino acid degradation, (c) in nucleic acid biosynthesis, and (d) in amino acid biosynthesis. The enzymes are indisputably homologous; even a simple sequence search identifies significant similarities. The colors show the analogy between the three catalyzed reactions from the perspective of organic chemistry. The functions of the proteins, from their roles in pathways, are quite different. An annotation strategy that assumes homologous proteins confer fitness in their host organisms in an analogous way would be misleading by this example.

But should they? All of these proteins use fumarate as a substrate. They all, in the language of the chemist, add the elements of H-X to fumarate using a Michael reaction, where the carboxylic acid functional group acts as an electron sink. This type of language is very close to that used by the Enzyme Commission when it assigns "EC" numbers to enzymes. In the language of the chemist, all of these proteins have analogous function because they all catalyze an E2 addition reaction to fumarate. Evolutionary recruitment in this family presumably occurred because of this mechanistic similarity (Gerlt and Babbitt, 1998).

The point to be made here is not that one cannot infer function by homology alone. Nor do we wish to argue that the biologist's view of function is right, while the Enzyme Commission's view is wrong. Rather, the point to be taken is that the analysis of function is tied to the language used to describe it. The language used to describe the systems determines whether one sees "equivalency" or "non-equivalency".

Orthologs as Functional Analogs?

Some attempts to alleviate these problems are based on the identification of orthologous sequences. Here, the homology-implies-equivalency assumption is restricted to a subset of homologs that diverged in the most recent common ancestor of the species sharing the homologs. This strategy is useful, of course. But it is likely to be far less general than is widely thought. Two species living in the same space, almost by axiom, cannot have identical strategies for survival. This, in turn, implies that two orthologous proteins may not contribute to fitness in *exactly* the same way in two species.

Some examples are useful. Leptin, for example, is known from genetics to be related to the obesity phenotype in the mouse. The human homolog, almost certainly the ortholog, is known, and is a target for drug development as an obesity gene in humans. Some details of the molecular history, however, suggested that it might not be. A reconstruction of the evolutionary history of the leptin family (Figure 5-2) shows that as primates emerged from the cenancestor of mouse and human, the leptin gene underwent an episode of rapid

sequence evolution involving many non-synonymous substitutions in the leptin gene (Benner *et al.*, 1998). Indeed, the reconstructed evolutionary history (Messier and Stewart, 1997) of the gene family shows that the number of nonsynonymous changes that accumulated in the gene during this episode, divided by the number of synonymous changes, normalized for the number of nonsynonymous and synonymous sites (the K_a/K_s ratio, sometimes referred to as ω , or dN/dS) is remarkably high. In fact, the K_a/K_s ratio in this episode is higher than that displayed by a pseudogene.

The only explanation consistent with Darwinian theory for this episode is that leptin was under "positive selection pressure" (Yang and Bielawski, 2000) as it entered the primate lineage 100 million years ago. Mutant forms of the primitive primate leptin evidently contributed more to the fitness of the primate descendants than non-mutant forms of the protein. This suggested, four years ago, that human "leptin" might not play a role in humans analogous to the role it plays in mice. At the very least, a primate model is recommended for pharmacological analysis of compounds targeted towards this system. And now, articles are appearing with titles such as "Whatever happened to leptin?" (Chircurel, 2000), noting that "the hormone's precise physical role seems to vary from species to species."

Analogous statements can be made about other pairs of orthologs from mammalian species (Chandrasekharan *et al.*, 1996; Liberles, *et al.*, 2001). Just as we cannot confidently accept annotation made by homology, we cannot be confident that annotations based on orthology are correct either.

In fact, "analogy", not "equivalency" nor "non-equivalency", is the topic in these examples of annotation. Analogy involves selection of some features of a system as being more important than others, and using these features to make a comparison. The Enzyme Commission views the structure of the substrate (fumarate) and the nature of the reaction being catalyzed (E2 addition, for example) as the features worth noting. The biologist (at least as represented above) considers the pathway as the noteworthy feature. The former is

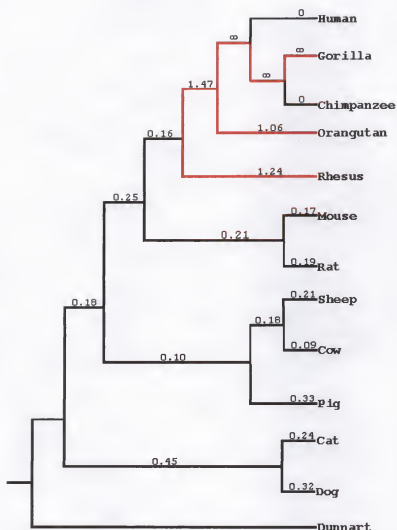


Figure 5-2. Evolutionary tree for leptins extracted from the Master Catalog. Numbers on the branches are K_a/K_s ratios, the ratio of nonsynonymous to synonymous changes, normalized for the number of nonsynonymous and synonymous sites in the gene. Undefined (∞) means that no silent substitutions occurred on the branch; calculating a K_a/K_s ratio would require division by zero. Reconstructed evolutionary sequences show rapid evolution of the leptin gene in primitive primates, consistent only with "positive selection" (where the lines are red), and implying a different "function" in primate leptins than in the cenancestral leptins (and rodent leptins). The branch with the K_a/K_s ratio of 1.47 (leading to the apes) contains 7.63 and 2.57 nonsynonymous and synonymous substitutions, respectively (before normalization), meaning that the high ratio is quite significant. The branch with the K_a/K_s ratio of 1.24 (leading to the rhesus monkey) contains 7.68 and 3.61 nonsynonymous and synonymous substitutions. The branch with the K_a/K_s ratio of 0.21 (leading to the rat/mouse ancestor) contains 14.31 and 31.62 nonsynonymous and synonymous substitutions.

more likely to be predictable from the molecule formula derived from the genomic sequence. The latter is closer to the Darwinian concept of fitness.

Again, neither view is "right". But the Wittgenstinian view of functional genomics requires that we understand the process and language of "analogy", recognize that it is not the same as "equivalency", and appreciate that an analogy is frequently more informative about the culture of the individual drawing the analogy than it is about the systems between which the analogy is being constructed.

A Behavioral/Functional Continuum

We can expect, almost from first principles, that the near continuum in molecular structure available to protein sequences is associated with a near-continuum of molecular behavior (Hey, 1999). This in turn, should be associated with a near continuum in fitness. Within this continuum, the case can be frequently made that the differences are more interesting than the similarities, and need to be captured and understood to make a useful functional annotation.

Consider, for example, the family of elongation factors (EFs) represented by EF-Tu (in bacteria) and EF-1 α (or eEF1A, in eukaryotes). All are annotated in the contemporary databases as having "the same function". After all, they all present a charged aminoacyl-tRNA to the ribosome. Closer inspection (Lockhart *et al.*, 1998; Moreira *et al.*, 1999; Gaucher *et al.*, 2001) shows, however, that the details by which this presentation is done, and the behaviors of individual EFs in general, are different, in a way that has an impact on any linguistic description of "function" (Figure 5-3). For example, EF may function in eukaryotes by binding to uncharged tRNAs in the nucleus, being charged there, and then being transporting to the cytosol via binding to actin. Regardless of the ability of bacteria EFs to display these behaviors (this is under-examined), the function of EFs in bacteria cannot contain this language, as bacteria do not have a nucleus.

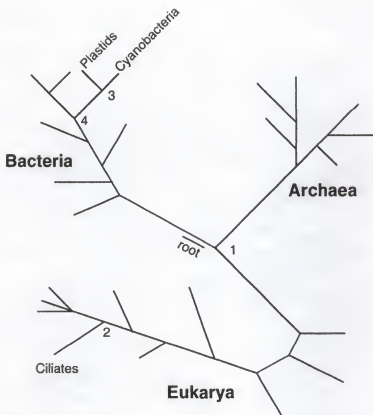


Figure 5-3. Continuum of elongation factor behavior. Recent studies demonstrate that the behaviors of various Elongation Factor Tu/1 α proteins are different in different members of the family, and these behavioral differences are functionally significant. Undoubtedly, "participation in translation" is the language describing one behavior almost certainly important to function (fitness) in all of these. Specific features of the behavior have, however, changed (even to the point of being gained or lost entirely) in the evolutionary episodes separating Nodes 1-4. Functional divergence in the GTP-, GDP-, tRNA-, and actin-binding domains has been demonstrated for the highly conserved EF protein. Shifts in EF behavior are found throughout the phylogenetic tree; between bacteria, archaea, and eukaryotes (Node 1), between ciliates and other eukaryotes (Node 2), between plastid and non-plastid bacteria (Node 3), and between photosynthetic and non-photosynthetic bacteria (Node 4). At the core of these studies lies the notion that functional importance is highly correlated with conserved evolutionary patterns. For example, ciliate EFs display functional divergence in the domains proposed to interact with actin. This is consistent with actin being a quantitatively minor protein and having an accelerated mutation rate in ciliates. A conventional homology-based search would have simply suggested ciliate EFs have "the same function" as other eukaryotic EFs due to their high sequence identity, and a substantial level of analogy in other functional behaviors. But it is also clear that a full understanding of a protein's function requires an analysis of the differences, and is best realized when sequences are placed within an historical, evolutionary comparative framework.

With EF's, the first level of annotation will undoubtedly reflect the analog in the functions of different proteins from different species. At the next level, however, the annotation must capture the differences. With EF's, the signature of functional change can also be found in the sequences, when they are viewed with a sufficiently sophisticated evolutionary model (Gaucher *et al.*, 2001).

A Way Forward

How might evolutionary analyses be used to generate linguistic statements concerning function for genomic sequences? The completeness of an organism's genomic sequence offers one advantage; it permits us to say what is *not* present. Further, we can draw on classical descriptions of the history of life known from paleontology and geology to contrast with the molecular histories of protein families reconstructed from genomic sequence databases. Functional genomics must approach genomic sequences in a particular way to facility this process.

(a) Complete evolutionary models of a protein family (Benner *et al.*, 2000).

Reconstructed sequences of ancient proteins, intermediates in evolutionary history of a protein family, need to be added to evolutionary models that include a multiple sequence alignment and an evolutionary tree. These ancestral sequences increase the scope of functional inferences that can be made from reconstructed evolutionary biology.

(b) Higher order analyses of sequence evolution. Today, the evolution of protein sequences is modeled using simple stochastic mathematics that treats proteins as if they were formless, functionless strings of letters. These models are poor approximations for reality. Their use comes from their ability to provide a "null hypothesis". The differences in how real proteins divergently evolve and how the stochastic models expect them to

evolve produces a signal, informative about form and function. Higher order analyses (Thorne *et al.*, 1992) of sequence divergence capture this signal. These incorporate substitution rates that depend on the site (gamma distribution) (Yang, 1996), non-independence of substitutions at different sites (Olmea *et al.*, 1999), and higher order gap penalties (Benner *et al.*, 1993). Many examples are now available where these higher order models support higher levels of sequence interpretation. Site-specific mutation rates are correlated to functionally important sites on a protein (Gaucher *et al.*, 2001). Shifts in functional constraints are evident when a specific site is rapidly evolving in one lineage but slowly evolving in another (covarion behavior) (Miyamoto and Fitch, 1995). Non-independence of sites is used for protein structure prediction (Benner *et al.*, 1997). We expect these types of analyses to be done routinely in the future (Jermann *et al.*, 1997; Golding and Dean, 1998; Naylor and Gerstein, 2000; Yang and Bielawski, 2000).

(c) Improve the dating of events in the reconstructed molecular history of the protein family. Genomics becomes especially powerful when events in the reconstructed molecular record are correlated with events in the geological and paleontological records. To make this correlation requires, however, a molecular clock. Amino acid sequences themselves are known to be imprecise molecular clocks (Ayala, 1999). Metrics that use synonymous substitutions are frequently used to date molecular events. We expect to see new tools that reflect the complexities of the mutation process to make dating more reliable, especially within vertebrate evolution (Peltier *et al.*, 2000). For example, the organic chemistry and selective mechanisms governing mutations within GC isochores can lead to spurious estimations when performing phylogenetic analyses. Incorporating more complex evolutionary models that account for biases in synonymous substitution

rates greatly enhance comparative analyses (Moore and Holmes, 2000). This, in turn, will open a new avenue for extracting information about function in an organismic and ecological context.

(d) Interpret sequence evolution within the context of three-dimensional structures.

The three dimensional structure of the protein connects sequence to reactivity.

Permutations within primary sequences can be correlated to those sites that are responsible for protein-ligand interactions and therefore differences in behavior. A three dimensional structure therefore adds significantly to any story in molecular evolution, and does so especially when complex phenomena are being analyzed (Miyamoto and Fitch, 1995; Golding and Dean, 1998; Gaucher *et al.*, 2001).

(e) Naturally structured protein sequence databases. After all of the genomes of all of the organisms on Earth are sequenced, all of the protein sequences almost certainly will be recognizable as being members of one of fewer than 10^5 protein families. A naturally structured database reflects this fact, organizing sequences according to their natural history. This organizational principle is exploited by Hovergen (Duret *et al.*, 1994), COG (Tatusov *et al.*, 1997), DMO (Gracy and Argos, 1998), Pfam (Bateman *et al.*, 2000), and the Master Catalog (Benner *et al.*, 2000).

Stories that combine part or all of these prescriptions are now emerging for many specific cases. These include: Myc and transferrins (Gu, 1999), elongation factors (Lockhart *et al.*, 1998; Moreira *et al.*, 1999; Gaucher *et al.*, 2001), ribonucleases (Jermann *et al.*, 1995), opsins (Chang and Donoghue, 2000), globins (Naylor and Gerstein, 2000), and lysozymes (Messier and Stewart, 1997). The ultimate goal, however, will be to join these specific cases into a unified model that combines the molecular history of life on

Earth with the record from natural history (Benner, 2001). Such a large-scale analysis will incorporate dates in the past, places on the globe, and events in the molecular geological and paleontological records, in a way that connects genes and proteins, their host organisms, and their ecosystems set in a planetary context.

APPENDIX MULTIPLE SEQUENCE ALIGNMENT OF ELONGATION FACTORS ANALYZED IN CHAPTER 1

	10	20	30	40	50	60
AGRTU						
ANANI	AKSKFERNKPHVNIGTIGHVDHGKTS	LTAAITKYF-----	GEFKAYDQIDAAPEEK			
AQUPY	ARAKFERTKPHANIGTIGHVDHGKTT	LTAAITTVLAKA----	GMAKARAYADIDAAPEEK			
BACFR	AKEKFERTKEHVNVGTIGHVDHGKST	LTSAITCVLAAGLVEGGKAKCFKYE	IDKAAPEEK			
BACST	AKEKFERTKPHVNIGTIGHVDHGKTT	LTAAITTVLAKQ----	GKAEAKAYDQIDAAPEER			
BACSU	AKEKFDRSKSHANIGTIGHVDHGKTT	LTAAITTVLHKK----	SGKGTAMAYDQIDGAPEER			
BRELN	AKASFERTKPHVNIGTIGHVDHGKTT	LTAAITKVLADQ---	YPDLNEARAFDQVDNAPEEK			
BURCE	AKGKFERTKPHVNVTIGHVDHGKTT	LTAAITTVLTCK----	FGGEAKAYDQIDAAPEEK			
CAMJE	AKEKFSRNKPHVNIGTIGHVDHGKTT	LTAAISAVLSRR----	GLAELKDYDNIDNAPEEK			
CORGL	AKAKFERTKPHVNIGTIGHVDHGKTT	TAAITKVLAADT---	YPELNEAFAFDSIDKAAPEEK			
CYFLY	AKETFDRSKPHLNIPTIGHVDHGKTT	LTAAITTVLANA----	GLSELRSFDSIDNAPEEK			
DEISP	AKGTERTKPHVNVGTIGHVDHGKTT	LTAAITFTAAS---	DPTIEKLAYDQIDKAAPEEK			
FERIS	AKVTFTVRKPHMNVGTIGQIDHGKTT	LTAITKYCSFF----	GWADYTPYEMIDKAAPEEK			
FLAFE	AKETFKREKPHVNIGTIGHVDHGKTT	LTAAITDILSKK----	GLAQAKKYDEIDGAPEEK			
MICLU	AKAKFERTKAHVNIPTIGHVDHGKTT	LTAATSKVLYDK---	YPDLNEARDFATIDSAPEEK			
MYCGA	AKERFDRSKPHVNIGTIGHIDHGKTT	LTAICTVLSKA----	GTSEAKKYDEIDAAPEEK			
MYCHO	AKLDFDRSKPHVNIGTIGHVDHGKTT	LTAATVLAQK----	GLAEARDYASIDNAPEEK			
MYCLE	AKAKFERTKPHVNIGTIGHVDHGKTT	LTAATKVLHDK---	FPNLNESRAFQIDNAPEEK			
NEIGO	AKEKFERSKPHVNVTIGHVDHGKTT	LTAALTILAKK----	FGGAAKAYDQIDNAPEEK			
FLARO	AAKAKLERTKPHMNIPTIGHIDHGKTT	LTAITKVLHNR---	YPELNKATPFDDKIDKAAPEEK			
FLAROB	AKAKFERTKPHMNIPTIGHIDHGKTT	LTAITKVLHNR---	YPELNKATPFDDKIDKAAPEEK			
RICPR	AKAKFERTKPHVNIGTIGHVDHGKTT	LTAAITILAKT----	GGAKATAYDQIDAAPEEK			
SALTY	SKEKFERTKPHVNVGTIGHVDHGKTT	LTAAITTVLAKT----	YGGAAARAFDQIDNAPEEK			
SHEPU	AKAKFERIKPHVNVTIGHVDHGKTT	LTAATSHVLAKT----	YGGAEKDFSDQIDNAPEEK			
SPIPL	ARAKFERNKPHVNIGTIGHVDHGKTT	LTAITMTLAAS----	GGAKARKYDDIDAAPEEK			
STIAU	AKEKFERNKPHVNIGTIGHVDHGKTT	LTAITKVLAKT----	GGATFLAYDQIDKAAPEEK			
STRAU	AKAKFERTKPHVNIGTIGHIDHGKTT	LTAATKVLHDK---	YPDLNAASAFDQIDKAAPEEK			
STRCJ	AKAKFERTKPHVNIGTIGHVDHGKTT	LTAATKVLHDA---	IPDLNFPFDFEIDKAAPEEK			
STROR	AKEYYDRSKPHVNIGTIGHVDHGKTT	LTAITTVLARR---	LPSAVNQPKDYASIDAAPEEK			
TAXOC	AKETFDRSKPHVNIGTIGHVDHGKTT	LTAAITTVLANK----	GLAAKDRFSSIDNAPEEK			
THEAQ	AKGEFIRKPHVNVGTIGHVDHGKTT	LTAALTYVAAA-----	NPNEVEKYDGDIDKAAPEEK			
THEMA	AKEKFVRTKPHVNVTIGHIDHGKST	LTAITKYLKSLK----	VLAQYIPYDQIDKAAPEEK			
THETH	AKGEFVRTKPHVNVGTIGHVDHGKTT	LTAALTYVAAA-----	NPNEVEKYDGDIDKAAPEEK			
THICU	AKSKFERTKPHVNVGTIGHVDHGKTT	LTAITTVLSSK----	FGGEAKAYDQIDNAPEEK			
TREHY	AKGTVEGNKTHVNVGTIGHVDHGKTT	LTSAITAVSSAM---	FPATVQKVAYDSVAKASESQ			
UREUR	AKAKFERTKPHVNIGTIGHVDHGKTT	LTAISTVLAKK----	GQAIAQSYADVDKTPPEER			
WOLSU	AKKGFVYKPHVNIGTIGHVDHGKTT	LTAASAVLATK----	GLCELKDYDAIDNAPEEK			
AQAAE1	AKSKFERTKEHVNVGTIGHVDHGKST	LTSAITCVLAAGLVEGGKAKCFKYE	IDKAAPEEK			
AQAAE2	AKEKFERTKEHVNVGTIGHVDHGKST	LTSAITCVLAAGLVEGGKAKCFKYE	IDKAAPEEK			
BBUR	AKEVFQRTKPHMNVGTIGHVDHGKTT	LTAATISYCSKL----	NKDAKALKYEDIDNAPEEK			
ECOL12	SKEKFERTKPHVNVGTIGHVDHGKTT	LTAITTVLAKT----	YGGAAARAFDQIDNAPEEK			
ECOL11	SKEKFERTKPHVNVGTIGHVDHGKTT	LTAITTVLAKT----	YGGAAARAFDQIDNAPEEK			
HINF	SKEKFERTKPHVNVGTIGHVDHGKTT	LTAITTVLAKH----	YGGAAARAFDQIDNAPEEK			
HPYL	AKEFNRKPHVNIGTIGHVDHGKTT	LTAASAVLSLK----	GLAEMKDYDNIDNAPEEK			

MGEN AREKFDRSKPHVNVGTIGHIDHGKTTLTAICTVLAKE----GKSAATRYDEIDKAPEEK
 MPNEU AREKFDRSKPHVNVGTIGHIDHGKTTLTAICTVLAKE----GKSAATRYDQIDKAPEEK
 MTUB AKAKFQRTKPHVNI GTIGHVDHGKTTLTAAITKVLHDK---FPDLNETKAFQIDNAPEER
 SYN3P ARAKFERTKDHVNI GTIGHVDHGKTTLTAAITMTLAE-----GGAKARKYEDIDAAPEEK
 SYN7P ARAKFERTKPHANI GTIGHVDHGKTTLTAAITTVLAKA-----GMAKARAYADIDAAPEEK
 TPAL AKEKFARTKVHMNVGTIGHVDHGKTTLSAAITSYCAKK----FGDQQLKYDEIDNAPEEK
 GLUPL GGGGGAEKPLLNVCFIGHVDSGKSTTVGNLAFQLGAIKMDKLKKEAEERGMDMSAAER
 METVA GGGGGAKTKPLLNVAFIGHVDAKGSTTVGRLLLDGGAILVLRLEKAEKKGMDGLEKER
 SULAC GGGGGGSKPHNLNLIVIGHVDHGKSTTLGRLLMDRGFITVKEAEAEAKKLGMDRLEKER

	70	80	90	100	110	120
AGRTU	ARGITISTAHVEYETPARHYAHVDCPGHADYVKNMITGAAEMDGA	ILVCSAADGMPQTR				
ANANI	ARGITINTAHVEYETGNRHYAHVDCPGHADYVKNMITGAAQMDGAI	LVVSAADGMPQTR				
AQUPY	ERGITINTHVEYETAKRHYAHVDCPGHADYIKNMITGAAQMDGAILV	VSAADGMPQTR				
BACFR	ERGITINTSHVEYETANRHYAHVDCPGHADYVKNMVTGAAQMDGAI	LVVAATDGPMPQTR				
BACST	ERGITISTAHVEYETEARHYAHVDCPGHADYVKNMITGAAQMDGAI	LVVSAADGMPQTR				
BACSU	ERGITISTAHVEYETETRHYAHVDCPGHADYVKNMITGAAQMDGAI	LVVSAADGMPQTR				
BRELN	ERGITINVSHEVYQTEKRHYAHVDAPGHADYVKNMITGAAQMDGAI	LVVAATDGPMPQTR				
BURCE	ARGITINTAHVEYETANRHYAHVDCPGHADYVKNMITGAAQMDGAILV	CSAADGMPQTR				
CAMJE	ERGITIATSHIEYETDNRHYAHVDCPGHADYVKNMITGAAQMDGAI	LVVSAADGMPQTR				
CORGL	ERGITINISHEVYQTEKRHYAHVDAPGHADYIKNMITGAAQMDGAI	LVVAATDGPMPQTR				
CYTLY	ERGITINTSHVEYSTANRHYAHVDCPGHADYVKNMVTGAAQMDGAILV	VAAATDGPMPQTR				
DEISP	ARGITINTAHVEYENTPRHYSHVDCPGHADYVKNMITGAAQMDGAILV	VSSADGMPQTR				
FERIS	ERGITINTHVEYQTEKRHYAHIDCPGHADYIKNMITGAAQMDGAILV	LAATDGPMPQTR				
FLAFE	ERGITINTAHVEYETANRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VASDGPMPQTR				
MICLU	QRGITINISHEVYQTEKRHYAHVDAPGHADYIKNMITGAAQMDGAILV	VAAATDGPMPQTR				
MYCGA	ARGITINTSHVEYATQNRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VASTDGPMPQTR				
MYCHO	ARGITINTSHIEYQTEKRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATDGPMPQTR				
MYCLE	QRGITINISHEVYQTEKRHYAHVDAPGHADYIKNMITGAAQMDGAILV	VAAATDGPMPQTR				
NEIGO	ARGITINTSHVEYETETRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				
PLAROA	ARGITISIAHVEYQTEKRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATDGPMPQTR				
PLAROB	ARGITISIAHVEYQTEKRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATAGMPQTR				
RICPR	ERGITISTAHVEYETQNRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				
SALTY	ARGITINTSHVEYDTPRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATDGPMPQTR				
SHEPU	ERGITINTSHIEYDTPSRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VASTDGPMPQTR				
SPIPL	QRGITINTAHVEYETEQRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				
STAU	ERGITISTAHVEYQTKNRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				
STRAU	QRGITISTAHVEYQTEARHYAHVDCPGHADYIKNMITGAAQMDGAILV	VAAATDGPMPQTR				
STRCJ	QRGITISTAHVEYQTESRHYAHVDCPGHADYIKNMITGAAQMDGAILV	VAAATDGPMPQTR				
STROK	ERGITINTAHVEYETEKRHYAHIDAPGHADYVKNMITGAAQMDGAILV	VASTDGPMPQTR				
TAXOC	ERGITINTAHVEYSTANRHYAHVDCPGHADYVKNMVTGAAQMDGAILV	VAAATDGPMPQTR				
THEAQ	ARGITINTAHVEYETAKRHYSHVDCPGHADYIKNMITGAAQMDGAILV	VSAADGMPQTR				
THEMA	ARGITINTHVEYQTEKRHYAHIDCPGHADYIKNMITGAAQMDGAILV	VAAATDGPMPQTR				
THETH	ARGITINTAHVEYETAKRHYSHVDCPGHADYIKNMITGAAQMDGAILV	VSAADGMPQTR				
THICU	ARGITINTAHVEYETANRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				
TREHY	GRLLTIATSHVEYEDNRHYAHVDCPGHADYIKNMITGAAQMDGAILV	VSAADGMPQTR				
UREUR	ERGITINASHVEYETKTRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VASDGVMAQTK				
WOLSU	ERGITIATSHIEYETENRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				
AQUAE1	ERGITINTHVEYETAKRHYAHVDCPGHADYIKNMITGAAQMDGAILV	VSAADGMPQTR				
AQUAE2	ERGITINTHVEYETAKRHYAHVDCPGHADYIKNMITGAAQMDGAILV	VSAADGMPQTR				
BBUR	ARGITINARIHIEYETANRHYAHVDCPGHADYIKNMITGAAQMDGAILV	VAAATDGPMPQTR				
ECOLI2	ARGITINTSHVEYDTPRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATDGPMPQTR				
ECOLI1	ARGITINTSHVEYDTPRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATDGPMPQTR				
HINF	ARGITINTSHVEYDTPRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VAAATDGPMPQTR				
HPYL	ERGITIATSHIEYETENRHYAHVDCPGHADYVKNMITGAAQMDGAILV	VSAADGMPQTR				

MGEN ARGITINSAHVEYSSDKRHYAHVDCPGHADYIKNMITGAAQMDGAILVVSATDSVMPQTR
 MPNEU ARGITINSAHVEYSSDKRHYAHVDCPGHADYIKNMITGAAQMDGAILVVSATDSVMPQTR
 MTUB QRGITINTIAHVEYQTDKRHYAHVADPGHADYIKNMITGAAQMDGAILVVAATDGPMPQTR
 SYN3P ARGITINTIAHVEYETDSRRHYAHVDCPGHADYVKNMITGAAQMDGAILVVSAAADGPMQTR
 SYN7P ARGITINTIAHVEYETGNRHYAHVDCPGHADYVKNMITGAAQMDGAILVVSAAADGPMQTR
 TPAL ARGITINTIRHLEYQSDRRHYAHIDCPGHADYVKNMITGAAQMDGAILVVSAPDGVMPQTK
 GLUPL ERGITITITSLMKLETSKHLNVIDCPGHQDFTKNNMTVGAAQADVGVVLVPCASCSITGLK
 METVA ERGVTIDVAHKKFTTAKYEVTIVDCPHRDEFTKNNMTGASQADAALVNVNDSGIVQTR
 SULAC ERGVTINLSFMRFTETRYFFTVIDAPGHRDFVKNMITGASQADAALVVSAKAGMSAQTR

	130	140	150	160	170	180	
AGRTU	EHILLARQVGVP	AIIVVFLNKVDQ	VDDAELELLE	VELEVR	RELLSSYDF	PFGDDIPIIKGSALA	
ANANI	EHILLAKQVGVP	NIIVVFLNKED	MVDDAELEL	VELEVR	RELLSSYDF	PFGDDIPIVAGSALQ	
AQUYP	EHVLLARQVNV	PYIVVFMNKC	DMVDDAELE	LELEVR	RELLSKYEF	PGDEVPIRGSALG	
BACFR	EHILLARQVNV	PKLVVFMNKC	DMVEDAELE	LEVEME	RELLSFYDF	DGNTPIIQGSALG	
BACST	EHILLSRQVG	VPYIVVFLNK	CDMVDDAE	LELEVE	MVRD	LLSEYDFPGDEVPIKGSALK	
BACSU	EHILLSKNVG	VPYIVVFLNK	CDMVDDAE	LELEVE	MVRD	LLSEYDFPGDDVPVVKGSALK	
BRELN	EHVLLARQVG	VPYIVVFLNK	SDMVDDAE	LELEVE	FVRD	LLSQDFDGNAPVPIVSALK	
BURCE	EHILLARQVG	VPYIIIVFLN	KCDSDVDDA	ELLEVEME	VR	RELLSKYDFPGDDTPIVKGSALK	
CAMJE	EHILLSRQVG	VPYIVVFMN	KADMVDDA	ELLEVEME	I	RELLSSYDFPGDDTPIISGSALK	
CORGL	EHVLLARQVG	VPYIILVALN	KCDMVEDEE	IELVEME	VR	RELLSAEQDYD-EEAPIVHISALK	
CYTLY	EHILLGRQVG	IPRIIVVFLN	KVDMVDDA	ELLEVEME	VR	RELLSFYDFDGNPVPVSGSALK	
DEISP	EHILLARQVG	VPYIVVFMN	KVDMVDDA	ELLEVEME	VR	RELLSKYEFPGDDLPVVKGSALQ	
FERIS	EHVLLARQVNV	PAMIVFINK	VDMV-DPEL	VDLVEME	VR	RELLSKYEFPGDEVFVVRGSALK	
FLAFE	EHILLAAQVG	VPKMOVFLN	KVDLVDDE	ELLEVEME	VR	RELLTKRFGDNTPIIKGSATG	
MICLU	EHVLLARQVG	VPALLVALN	KSDMVEDEE	LELVEME	VR	RELLSFYDFDGNAPVPIVSALK	
MYCGA	EHILLARQVG	VPKMOVFLN	KCDVADDP	QELVEME	VR	RELLSKYDFGDDTPVIRGSALK	
MYCHO	EHILLARQVG	VPKIVVFLN	KIDMFKDD	EMVGLV	EM	DVRSLLSEYDFGDNAPPIIAGSALK	
MYCLE	EHVLLARQVG	VPYIILVALN	KSDA	VDDAELE	LEVEME	VR	RELLSAAEQFD-EDAPVVRVSALK
NEIGO	EHILLARQVG	VPYIIVFMN	KCDMVDDA	ELFQVEME	I	RELLSSYDFPGDDCPPIVQGSALK	
PLAROA	EHVLLARQVG	VPYIVVVALN	KADMVDDA	ELLEVEME	VR	RELLSAAEQFPFGDDLPVVRVSALK	
PLAROB	EHVLLARQVG	VPYIVVVALN	KADMVDDA	ELLEVEME	VR	RELLSAAEQFPFGDDLPVVRVSALK	
RICPR	EHILLAKQVG	VPAMVFLN	KVDMVDDP	DLLEVEME	VR	RELLSKYDFPGNEIPIIKGSALQ	
SALTY	EHILLGRQVG	VPYIIVFLN	KCDMVDDA	ELLEVEME	VR	RELLSQYDFPGDDTPVIRGSALK	
SHEPU	EHILLSRQVG	VPFFIIVFMN	KCDMVDDA	ELLEVEME	VR	RELLSSYDFPGDDLPVQGSALK	
SPIPL	EHILLAKQVG	VPISIVVFLN	KADMVDDA	ELLEVEME	VR	RELLSSYDFPGDDIPIVSGSALK	
STIAU	EHILLARQVG	VPYIVVFLN	KVMDLDP	PELREL	VEME	VR	RELLKKYEFPGDSTPIIPGSALK
STRAU	EHVLLARQVG	VPYIIVVALN	KADMVDDA	ELLEVEME	VR	RELLSSYDFPGDDLPVQVGSALK	
STRCJ	EHVLLARQVG	VPYIIVVALN	KADMVDDA	ELLEVEME	VR	RELLSEYDFDGNCPVQVGSALK	
STROH	EHILLSRQVG	VKHLIVFMN	KIDLVDDE	ELLEVEME	I	RELLSEYDFPGDDLPVQGSALK	
TAXOC	EHILLARQVG	VPQLVVFMN	KVDMVDDP	ELLEVEME	I	RELLSFYDFDGNIPVQVGSALK	
THEAQ	EHILLARQVG	VPYIVVFMN	KVDMVDDP	ELLDVEME	VR	RELLLNQYEFPGDEVPIRGSALK	
THEMA	EHVLLARQVG	VEPYMIVFINK	TDMDVDDP	ELIDLVEME	VR	RELLDSQYFPGDEVPIRGSALK	
THETH	EHILLARQVG	VPYIIVVFMN	KVDMVDDP	ELLDVEME	VR	RELLLNQYEFPGDEVPIRGSALK	
THICU	EHILLARQVG	VPYIIVFLN	KCDMVDDA	ELLEVEME	VR	RELLSKYDFPGDDTPIIKGSALK	
TREHY	EHVLLSRQVG	YNIVVFLN	KCDKLDPE	MAEIVEA	EDIVLD	HYFGDGSKTPPIRGSALK	
UREUR	EHILLARQVG	VPKIVVFLN	KCDFMTD	PDMDQLVEME	VR	RELLSKYDFGDDTPVIRGSALK	
WOLSU	EHILLSRQVG	VPYIVVFLN	KEDMVDDA	ELLEVEME	VR	RELLSNYDFPGDDTPIVAGSALK	
AQUAE1	EHVLLARQVNV	PYIIVVFMN	KCDMVDDA	ELLEVEME	VR	RELLSKYEFPGDEVPIRGSALK	
AQUAE2	EHVLLARQVNV	PYIIVVFMN	KCDMVDDA	ELLEVEME	VR	RELLSKYEFPGDEVPIRGSALK	
BBUR	EHILLAQRMG	IKKIIVFLN	KDLA-DPEL	VELVEVE	VELV	EKYGFS-ADTPIKGSALG	
ECOLI2	EHILLGRQVG	VPYIIVFLN	KCDMVDDA	ELLEVEME	VR	RELLSQYDFPGDDTPIVIRGSALK	
ECOLI1	EHILLGRQVG	VPYIIVFLN	KCDMVDDA	ELLEVEME	VR	RELLSQYDFPGDDTPIVIRGSALK	
HINF	EHILLGRQVG	VPYIIVFLN	KCDMVDDA	ELLEVEME	VR	RELLSQYDFPGDDTPIVIRGSALK	
HPYL	EHILLSRQVG	VPYIIVVFLN	KQDMVDDQ	ELLEVEME	VR	RELLSAYEFPGDDTPIVAGSALK	

MGEN EHILLARQVGVPMVFLNKC DIASDEEVQELVAEEVRDLLTSYGF DGKNTPIIYGSALK
 MPNEU EHILLARQVGVPMVFLNKC DIATDEEVQELVAEEVRDLLTSYGF DGKNTPIIYGSALK
 MTUB EHVLARQVGVPIYLVALNKADAVDDEELLELVEMEVRLLAAQEPD-EDAPVVRVSALK
 SYN3P EHILLAKQVGVPKLVFLNKKDMVDDEELLELVELEVRELLSYDFPGDDIPIVAGSALK
 SYN7P EHILLAKQVGVPIVFLNKE DMVDDAELELVELEVRELLSYDFPGDDIPIVAGSALK
 TPAL EHLLARQVGVPSIIVFLNKVDLDDPELLELVEEEVRDALAGYGF-RETPIVKGSAFK
 GLUPL DHIMISGVLGCRKLIIVCNKVDTIDEKNRISRFDEVAKEMKGI IAKSKDKPIIIPISGYL
 METVA EHVFILRTLGRQLAVAVNKMDTFSEADYNELKKMIGDQLLKMIFNPEQINFPVVASLH
 SULAC EHILSKTMGINQVIVAINKMDLYDEKRFKEIVDTIVSKFMKSGFGDMNKVKEFVVPAD

	190	200	210	220	230	240
AGRTU	ALEDSK-----	KIGEDAIRELMAAVDAYI	PTPERPIDQF	FLMPIEDVFS	ISGRGTV	
ANANI	ALEAIQGGASGQ	GDNPWVKILKLMEEVDAYI	PTPEREVRDP	FLMAVEDVFT	ITGRGTV	
AQUFY	ALQELQNSPGK----	WVGSIKELLNAMDEYI	PTPEREDVK	FLMPIEDVFS	ISGRGTV	
BACFR	ALNGVEK-----	WEDKVMELMAEVDTWI	PLPRDVK	KFLMPVEDVFS	ITGRGTV	
BACST	ALEGDPK-----	WEEKI IELMNAVDYI	PTPQREV	DKPFMMPIEDVFS	ITGRGTV	
BACSU	ALEGDAE-----	WEAKIFELMDAVDEYI	PTPERDTEK	PFMMPEVEDVFS	ITGRGTV	
BRELN	ALEGDEK-----	WVKSVDLMAAVDDNV	PEPERDVK	KFLMPVEDVFT	ITGRGTV	
BURCE	ALEGDTGE-----	LGEVAIMSLADALDTYI	PTPERAVDGA	FLMPVEDVFS	ISGRGTV	
CAMJE	ALEEAKAGQDGE----	WSAKIMDLMAAVDSYI	PTPTRDTEK	FLMPIEDVFS	ISGRGTV	
CORGL	ALEGDEK-----	WGQILELMQACDDNI	PDPVRETD	KFLMPIEDIFT	ITGRGTV	
CYTLY	ALNGEQK-----	WVDTVMELMAEVDNW	IELPKRDVK	DFLMPVEDVFT	ITGRGTV	
DEISP	ALEALQANPKTARGED	KWVDRIWELLDAVDSYI	PTPERATDK	FLMPVEDVFT	ITGRGTV	
FERIS	ATEAPNDNDP-----	AYKPIKELLDAMDYTF	PDPVREVDK	FLMPIEDVFS	ITGRGTV	
FLAFE	ALAGEEK-----	WVKE IENLMDAVDSYI	PLPRVDL	PFMSVEDVFS	ITGRGTV	
MICLU	ALEGDPQ-----	WVKSVEDLMDAVDEYI	PDPVRDVK	KFLMPIEDVFT	ITGRGTV	
MYGCA	ALNGEPA-----	WEEKI HELMKAVDEYI	PTPDREV	DKFLLPIEDTMT	ITGRGTV	
MYCHO	ALQGDPPE-----	YEGGILELMDAVDTYI	EEPKRET	DKFLLMAVEDVFT	ITGRGTV	
MYCLE	ALEGDAK-----	WVESVTQLMDAVDESI	PAPVRETD	KFLMPVEDVFT	ITGRGTV	
NEIGO	ALEGDA-----	YEEKIFELATALDRYI	PTPERAVDK	FLMPIEDVFS	ISGRGTV	
PLAROA	ALEGDEK-----	WADSI IELMNAVDENI	PEPRDVK	KFLMPIEDVFS	ITGRGTV	
PLAROB	ALEGDEK-----	WADSI IELMNAVDENI	PEPRDVK	KFLMPIEDVFS	ITGRGTV	
RICPR	ALEGKPE-----	GEKAINELMNAVDYI	PQPIELQDK	FLMPIEDVFS	ISGRGTV	
SALTY	ALEGDAE-----	WEAKI IELAGFLDSYI	PEPERAIDK	FLMPIEDVFS	ISGRGTV	
SHEPU	ALEGEPE-----	WEAKILELAAALDSYI	PEPQRIDK	FLMPIEDVFS	ISGRGTV	
SPIPL	ALDFTLENPKTTRGEN	DWDKIHALMDEVDAYI	PTPERDIDK	GLDGLDEVFS	ITGRGTV	
STIAU	ALEGDTS-----	DIGEGAILKMAAVDEYI	PTPQATDK	FLMPVEDVFS	ISAGRGTV	
STRAU	ALEGDK-----	WGDKLLGLMDAVDEAI	PTPERDVK	FLMPVEDVFT	ITGRGTV	
STRCJ	ALEGDK-----	WGEKLLGLMKAVDENI	PQPERDVK	KFLMPIEDVFT	ITGRGTV	
STROR	ALEGDSK-----	YEDI IMELMNTVDEYI	PEPERDTEK	FLLPVEDVFS	ITGRGTV	
TAXOC	GLNGDAK-----	WVGTEIQLMDSVDNW	PIPPRLD	KFLMPVEDVFS	ITGRGTV	
THEAQ	ALEEMHKNPKTKRGEN	EWWDKIWELLDAI	DEYI	PTPVRDVK	KFLMPVEDVFT	ITGRGTV
THEMA	AVEAPNDPN-----	HEAYKPIQELLDAMDNYI	PDPQRDVK	KFLMPIEDVFS	ITGRGTV	
THETH	ALEQMHRNPKTKRGEN	EWWDKIWELLDAI	DEYI	PTPVRDVK	FLMPVEDVFT	ITGRGTV
THICU	ALEGDKG-----	ELGEGAILKLAELDTYI	PTPERAVDGA	FLMPVEDVFS	ISGRGTV	
THREY	AIQAIEAGKDFR-----	TDPCCKILDLLNALDTYI	PDPVREVDK	FLMPIEDVYSI	ISGRGTV	
UREUR	ALEGDPV-----	WEAKI DELMDAVDSWI	PLPERSTDK	FLLAIEDVFT	ISGRGTV	
WOLSU	ALEEANDQENGV-----	EWGEKVLKMAEVDRIYI	PTPERDVK	KFLMPVEDVFS	ISAGRGTV	
AQUAE1	ALQELEQNSP-----	GKWVESIKELLNAMDEYI	PTPQREV	DKFLMPIEDVFS	ISGRGTV	
AQUAE2	ALQELEQNSP-----	GKWVESIKELLNAMDEYI	PTPQREV	DKFLMPIEDVFS	ISGRGTV	
BBUR	AMSNPD-----	PESTKCVKELLESMDNYI	FDLPERDIDK	FLLAVEDVFS	ISGRGTV	
ECOL12	ALEGDAE-----	WEAKILELAGFLDSYI	PEPERAIDK	FLLPPIEDVFS	ISGRGTV	
ECOL11	ALEGDAE-----	WEAKILELAGFLDSYI	PEPERAIDK	FLLPPIEDVFS	ISGRGTV	
HINF	ALNGVAE-----	WEEKILELANHLDYI	PEPERAIDQ	FLLPPIEDVFS	ISGRGTV	
HPYL	ALEEAKAGNVG-----	EWGEKVLKMAEVDAYI	PTPERDTEK	FLMPVEDVFS	ISAGRGTV	

MGEN ALEGDPK-----WEAKIHDLIKAVDEWIPTPTREVDKPFLLAIEDTMTITGRGTV
 MPNEU ALEGDPK-----WEAKIHDLMNAVDDEWIPTPEREVDKPFLLAIEDTMTITGRGTV
 MTUB ALEGDAK-----WVASVEELMNAVDDESI PDPVRETDKPFLLMPVEDVFTITGRGTV
 SYN3P ALEGEKE-----YKDAILELMKAVDDYIDTPEREVDKPFLLMAVEDVFTITGRGTV
 SYN7P ALEAIQGGASGQKGDNPWVDKILKLMEEVDAYIPTPEREVDKPFLLMAVEDVFTITGRGTV
 TPAL ALQDGAS-----PEDAACIEELLAAMDYSFEDPVRDDARPFLLSIEDVYITISGRGTV
 GLUPL GINIVEKDGKFE-----WFGWKTLLEGALNSQIPPPRPIDKPLMPIDSHKIPGIGMV
 METVA GDNVFKKSERNP-----WYKGPKTIAEVIDGFPPEKPTNLRPLRIQDVYITIGVGT
 SULAC GDNVTHKSTKMP-----WYNGPKTLEELLQLEIPPPKPVDKPLRIPIQEVYSISGVGV

	250	260	270	280	290	300
AGRTU	VTGRVERGIVKVGEEVEIVGIR	-PTS	SKTTVT	VGEMFRKLLDQ	QAGDNIGAL	VRGVTRDG
ANANI	ATGRIERGVS	SVKGETIE	IVGLR-D	TRSTTV	TGVMFQK	TLDDEGLAGDNVGLLLRGV
AQUY	VTGRVERGVLRPGDEVEI	VGRLREE	PLKT	VATSIEMFRKVL	DEALPGDNI	GVLLRGVKGDD
BACFR	ATGRIETGV	IHVGDIE	IELGLG-	EDKKS	VVTGVMFRKLLD	QGEAGDNVGLLLRGV
BACST	ATGRVERGTLKVGDPVEI	IGLSDE	PKATT	VTGVMFRKLLD	QAEAGDNIGALLRGV	SRDE
BACSU	ATGRVERGQVKVGDEVEI	IGLQEE	NKTT	VTGVMFRKLLDY	AEAGDNIGALLRGV	SREE
BRELN	VTGRVERGVLLPNDEI	IVGIEK	KS	SKTTVTAIEMFRK	TLPPDARAGEN	VGLLLRGTKRED
BURCE	VTGRVERGIVKVGEEI	IVGIK-	PTVKT	CTGVMFRKLLD	QQAGDNVGLLLRG	TKRED
CAMJE	VTGRIEKGVVKGDTIE	IVGIK-D	TQTT	VTGVMFRKEMD	QGEAGDNVGLLLRG	TKKEE
CORGL	VTGRVERGTLNVNDDV	DIIGI	KEK	STSTVTG	IEFRKLLDSAEAGDN	CGLLLRGIKRED
CYTLY	ATGRIETGVANTGDAVD	I	GMGAD	KLASTITGVMFRK	ILDRGEAGDNVGL	LRGIEKQ
DEISP	ATGRVERGVVKVQDEVEI	IGLR-D	TKKT	VTGVIEMHRKLLD	SGMAGDNVGL	LRGVARD
FERIS	VTGRIERGVIKPGVEAEI	IGMSYE	IKKT	VITSVEMFRKEL	DEAIAAGDNV	GCLLRGSSKDE
FLAFE	ATGRIERGRIKVGEPVEI	VGLOES	PLNST	VTGVMFRKLLD	GEAGDNAGLLRG	VEKQT
MICLU	VTGRAERGT	LKINSE	VEIVGIR-	DVQKT	VTGVIEMFRK	QLDEAWAGENCGLLRGL
MYCGA	VTGRVERGQLKVGEEVEI	VGIT-D	TRKVV	VTGVIEMFRKLLD	AAAGDNAGILLRG	VDRKD
MYCHO	ATGRVERGVQLNEEVEI	VGILK-P	TKKT	VVTGVIEMFRK	NLKEAQAGDNAG	LLLRGIDRSE
MYCLE	VTGRVERGVNVNNEEVEI	VGIRQ	TTT	KTVTGVMFRKLLD	QQAGDNVGLLLRG	IKRED
NEIGO	VTGRVERGIIHVGDIEI	VGILK-E	TQKT	CTGVMFRKLLD	GEAGDNVGLLLRG	TKRED
PLAROA	VTGRIERG	VVKVNEQVDI	IGIKSE	TTTTVT	SIEMFNKMLDE	HAGDNAALLRGIKRE
PLAROB	VTGRIERG	VVKVNEQVDI	IGIKSE	TTTTVT	SIEMFNKMLDE	HAGDNAALLRGIKRE
RICPR	VTGRVESGIIKVGEEI	IVGLK-N	TQKT	CTGVMFRKLLD	EQSGDNVGL	LRGTKREE
SALTY	VTGRVERGIIKVGEEVEI	VGIR-E	TQKT	CTGVMFRKLLD	DEGRAGEN	VNGVLLRGIKREE
SHEPU	VTGRVERGIVRVGDVEI	VGIR-AT	TKT	CTGVMFRKLLD	DEGRAGEN	CGLLRGTKRED
SPIPL	STAGIERGKVKVGDT	VELIGIK-D	TRTT	VTGAE	MFQKTL	LEEGMAGDNVGLLLRG
STIAU	ATGRVERGKIKVGEEVEI	VGIR-P	TQKT	VTGVMFRKLLD	EGMAGDNIGALLRG	LRKRED
STRAU	VTGRIERGVLKVN	ETVDIIGI	KEK	TTTTVTG	IEFRKLLD	GEAGDNVGLLLRG
STRCJ	VTGRIERGVLKVN	ETVDIIGI	KEK	TTTTVTG	IEFRKLLD	GEAGDNVGLLLRG
STORR	ASGRIDRG	TVRVNDEI	IVGIEE	TQKAV	VTGVMFRKQL	DEGLAGDNVGLLRGV
TAXOC	ATGRIERG	VINSGE	PVEIL	MGAE	NKLSVTGVMFRK	LLDRGEAGDNVGLLLRG
THEAQ	ATGRIERGKVKVGDEVEI	IVGLAP	ETRKT	VVTGVMFRK	TLQEG	IAGDNVGLLLRGV
THEMA	VTGRIERGIR	RPGDVEI	IGLSYE	IKKT	VVTSVEMFRKEL	DEGIAGDNVGLLRG
THETH	ATGRIERGKVKVGDEVEI	IVGLAP	ETRT	VVTGVMFRK	TLQEG	IAGDNVGLLRGV
THICU	VTGRVERGIIKVGEEI	IVGLK-PT	LKT	CTGVMFRKLLD	QQAGDNVGLLLRG	TKREE
TREHY	VTGRIERGKIEKNEVEI	VGIR-PT	QKT	CTGVMFRK	KEVV-GIAG	VNGVLLRGV
UREUR	VTGRVERGVLKVNDEVEI	VGILK-D	TQKT	VVTGVIEMFRK	SLDQAEAGDN	NAGILLRGIKKED
WOLSU	VTGRIERG	VVKVGDVEI	VGIR-N	TQKT	VTGVMFRKEL	DKGEAGDNVGLLRGT
AQUAE1	VTGRVERGVL	RPGDVEI	VGRLREE	PLKT	VATSIEMFRKVL	DEALPGDNI
AQUAE2	VTGRVERGVL	RPGDVEI	VGRLREE	PLKT	VATSIEMFRKVL	DEALPGDNI
BBUR	ATGRIERGIIKVGQVEI	VGIR-ET	RKT	VTGVMFQK	LEGGQAGDNVGL	LRGVGRDD
ECOLI2	VTGRVERGIIKVGEEVEI	VGIR-E	TQKT	CTGVMFRKLLD	DEGRAGEN	VNGVLLRGIKREE
ECOLI1	VTGRVERGIIKVGEEVEI	VGIR-E	TQKT	CTGVMFRKLLD	DEGRAGEN	VNGVLLRGIKREE
HINF	VTGRVERGI	IRTGDEVEI	VGIR-D	TAKT	VTGVMFRKLLD	DEGRAGENIGALLRG
HPYL	VTGRIERG	VVKVGDEVEI	VGIR-PT	QKT	VTGVMFRK	LEKGEAGDNVGLLRGT

MGNE VTGRVERGELKVGQVEIVGLK-PIRKAVVTGIEMFKKELDSAMAGDNAGVLLRGVERKE
 MPNEU VTGRVERGELKVGQVEIVGLR-PIRKAVVTGIEMFKKELDSAMAGDNAGVLLRGVDKKE
 MTUB VTGRVERGVINVNEEVEIVGIRPSTTKTTVTGVEMFRKLLDQOQAGDNVGLLLRGVKRED
 SYN3P ATGRIERGKVVVGGEEISVIGIK-DTRKATVTGVEMFQKTLIEEGMAGDNVGLLLRGIOKED
 SYN7P ATGRIERGKVVVGGETIEIVGLR-DTRSTVTGVEMFQKTLDEGLAGDNVGLLLRGIOKED
 TPAL VTGRIECGVISLNEEVEIVGIRK-PTKKTVVVTGIEMFKNLLDQGIAGDNVGLLLRGVDKKE
 GLUPL YTG RVSTGAIKPGMVEIVSSQPTGVVAEVKTLIEIKHSRAAVVSGENCQVALKAASQGN
 METVA PVGRVETGIIKPGDEKVVFEPPAGAIGETKTVEMHHEQLPSAEPGDNIGFNVRGKGDKE
 SULAC PVGRIESGVLKVGDEKIVFMPVVGKIGEVRSIETHHTKIDKAEPGDNIGFNVRGVEKKD

	310	320	330	340	350	360
AGRTU	VERGQILCKPGSVKPHKKFMAEAYILTKEEGGRHTPFFTNRYRQPFYFRITDVTG-IVSLP					
ANANI	IERGVMVLAKPGSITPHTKFESEVYVLSKEEGGRHTPFFGPGYRQPFYVRTDVTGAIISDFT					
AQUYF	VERGQVLAQPGSVKAHRRFRAQVYVLSKEEGGRHTPFFVNYRQPFYFRITADVTGTVVKLP					
BACFR	IKRGMVLCCKPGQIKPHSKFKAQVYVLSKEEGGRHTPFFHNYRQPFYFRITMDCTG-EITLP					
BACST	VERGQVLAKPGSITPHTKFKAQVYVLSKEEGGRHTPFFSNRYRQPFYFRITDVTG-IITLP					
BACSU	IQRGMVLAKPGTITPHSKFKAQVYVLSKEEGGRHTPFFSNRYRQPFYFRITDVTG-IIHLP					
BRELN	VERGQVIVKPGSITPHTKFEAQVYVLSKDEGGRHNPFFSNRYRQPFYFRITDVTG-VITLP					
BURCE	VERGQVLAKPGSITPHTHFTAQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-SIELP					
CAMJE	VIRGMVLAKPGSITPHTDFAQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-SIKLA					
CORGL	VERGQVIVKPGAYTPHTEFEGSVYVLSKDEGGRHTPFFDNRYRQPFYFRITDVTG-VVKLP					
CYTLY	ISRGMVICKPGSVKPHSKFEAQVYVLSKEEGGRHTPFFHNYRQPFYFRITDVTG-TISLP					
DEISP	VERGQVLAKPGSIPKHTKFEASVYVLSKDEGGRHSAFFGGYRQPFYFRITDVTG-VVLEP					
FERIS	VERGQVLAKPGSITPLKKFKANIYVLSKEEGGRHTPFTKGYPQPFYFRITADVTGEIVDLP					
FLAFE	IRRGMVIVKPGSITPHTDKGEVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-EVELN					
MICLU	VERGQVLVEPGSITPHTNFEANVYVLSKDEGGRHTPFFSNRYRQPFYFRITDVTG-VITLP					
MYCGA	VQRGMVLAKPGSITPHKKFRAEIALYVLSKDEGGRHTAFLNGYRQPFYFRITDVTG-SIQLK					
MYCHO	VERGQVLAKPGTIVPHTQFEATVYVLSKEEGGRHTPFFHNYRQPFYFRITDVTG-GIEFK					
MYCLE	VERGQVVIKPGTTTTPHTEFEGQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-VVTLP					
NEIGO	VERGQVLAKPGTITPHTKFKAQVYVLSKEEGGRHTPFFHNYRQPFYFRITDVTG-TITLE					
PLAROA	VERGQCI IKPGTTTTPHTEFQAQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-VVNLN					
PLAROB	VERGQCI IKPGTTTTPHTEFQAQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-VVNLN					
RICPR	VERGQVLAKPGSIPKHDKFEAQVYVLSKEEGGRHTPFTNDYRQPFYFRITDVTG-IITKL					
SALTY	IERGQVLAKPGTIPKHTKFESEVYVLSKDEGGRHTPFFGKYRQPFYFRITDVTG-TIELP					
SHEPU	VERGQVLAKPGSINPHTTFESEVYVLSKEEGGRHTPFFGKYRQPFYFRITDVTG-TIELP					
SPIPL	VQRGMVIAKPGSITPHTKFEAQVYVLSKEEGGRHTPFFGKYRQPFYFRITDVTGIDEFT					
STIAU	LERGQVLANWGSINPHTKFAQVYVLSKEEGGRHTPFFGKYRQPFYFRITDVTG-TVKLP					
STRAU	VERGQVVIKPGSVTPHTEFQAQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-VVTLN					
STRJC	VERGQCI IKPGTIVTPHTEFATAYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-VVTLK					
STROR	IERGQVIAKPGSINPHTKFKGEVYILTKEEGGRHTPFFNRYRQPFYFRITDVTG-SIELP					
TAXOC	IRRGMVICKPGSVTPHKKFKAQVYVLSKEEGGRHTPFFNRYRQPFYFRITDVTG-IISLA					
THEAQ	VERGQVLAKPGSITPHTKFEASVYVLSKEEGGRHTGFFGYRQPFYFRITDVTG-VVRLP					
THEMA	VERGQVLAQPGSIPKPHKFKAQIYVLSKEEGGRHTPFTGKYRQPFYFRITADVTGIEIVGLP					
THETH	VERGQVLAKPGSITPHTKFEASVYVLSKEEGGRHTGFFSGYRQPFYFRITDVTG-VVQLP					
THICU	VERGQVLCKPGSIPKHTHTFAQVYVLSKDEGGRHTPFFNRYRQPFYFRITDVTG-AIELP					
TREHY	VERGQVLAKPGTITPHKKFKAQVYVLSKEEGGRHTPFFGKYRQPFYFRITDVTG-VINLQ					
UREUR	VERGQVLCKPGSIPKPHRTFAKVIYVLSKEEGGRHTPIVSGYRQPFYFRITDVTG-AISLP					
WOLSU	VERGMVLCKGSIPTHTNFEQEVYVLSKEEGGRHTPFFNGYRQPFYFRITDVTG-SISLP					
AQUAE1	VERGQVLAQPGSVKAHRRFRAQVYVLSKEEGGRHTPFFVNYRQPFYFRITADVTGTVVKLP					
AQUAE2	VERGQVLAQPGSVKAHRRFRAQVYVLSKEEGGRHTPFFVNYRQPFYFRITADVTGTVVKLP					
BBUR	IERGQVLSAPGTITPHKKFKASICYLTKKEEGGRHTPFFGKYRQPFYFRITDVTG-VVAL-					
ECOLI2	IERGQVLAKPGTIPKHTKFESEVYVLSKDEGGRHTPFFGKYRQPFYFRITDVTG-TIELP					
ECOLI1	IERGQVLAKPGTIPKHTKFESEVYVLSKDEGGRHTPFFGKYRQPFYFRITDVTG-TIELP					
HINF	IERGQVLAKPGSITPHTDFAQVYVLSKDEGGRHTPFFGKYRQPFYFRITDVTG-TIELP					
HPYL	VERGMVLCKPGSITPHKKFEGEIVYVLSKEEGGRHTPFFTNRYRQPFYFRITDVTG-SITLP					

MGEN VERGQVLAKPGSIKPHKKFKAEIYALKKEEGGRHTGFLNGYRQFYFRTTDTVGTG-SIALA
 MPNEU VERGQVLAKPGSIKPHKKFKAEIYALKKEEGGRHTGFLNGYRQFYFRTTDTVGTG-SISLP
 MTUB VERGQVVTLPKGTTPPHTEFEGQVYILSKDEGGRHTPPFNMYRQFYFRTTDTVGTG-VVTLP
 SYN3P IERGMVLAKPGSITPHTEFEGEVYVLKKEEGGRHTPPFNMYRQFYFRTTDTVGTIKSYT
 SYN7P IERGMVLAKPGSITPHTKFESEVYVLKKEEGGRHTPPFPQYRQFYFRTTDTVGTGAIISDFT
 TPAL VERGQVLSKPGSIKPHTKFEAQIYVLSKEEGGRHSPPFQGYRQFYFRTTDTVGTG-TISLP
 GLUPL IKPGHVFSNTKDVEI FEAARAKIVVAHPPKPGYCPTMDLGHVPCQITKFIS-KRMPG
 METVA IKRGDVLGHTTNPVTATDFTAQIVVLQHPSTDGTPVFVHTHTAQIACTFAEIQK-LNPAT
 SULAC VKRGDVGASVQNPTVADEFTAQVIVIHWPITGVGYTPVLHVHTASIAICRVSEITS-IDPKT

	370	380	390	400	409
AGRTU	EGTEMVMPGDNVTVEVELI	VP	IA	ME	EKLRFAIREGGRTVGAGIVASIVE
ANANI	ADDGMV	IPGDR	IKMTVELIN	PIA	IEQGMRFaireGGRTIGAGVVS
AQUPY	EGVEMVMPGDNVLE	LEVELI	APV	ALE	EGLRFAIREGGRTVGAGVVTKILD
BACFR	EGTEMVMPGDNVTIT	VELI	YP	PVAL	NIGLRFaireGGRTVGAGQITEIID
BACST	EGVEMVMPGDNVEM	VELI	APV	IA	IEEGTKFSIREGGRTVGAGSVSEIIE
BACSU	EGVEMVMPGDNTEM	NVELI	STIA	IE	EEGRFISIREGGRTVGSQVSTITE
BRELN	EGTEMVMPGDN	TMSVELI	QPI	AM	EDRLRFaireGGRTVGAGRVTKITA
BURCE	KDKEMVMPGDNV	SITVKLI	APV	IA	MEEGLRFAIREGGRTVGAGVVAKILD
CAMJE	DGVEMVMPGEN	VRI	TVSLI	APV	ALEEGRFAIREGGRTVGSQVVS
CORGL	EGTEMVMPGDNV	MSVTLI	QPV	AM	DEGLRFAIREGSR
CYTLY	SGVEMVMPGDNLT	ITVELLS	PI	AL	SEGLRFAIREGGRTVGAGQVTKIIE
DEISP	EGVEMVMPGDNIT	FVVELI	KPI	IA	MEEGLRFAIREGGRTVGAGVVAKVLE
FERIS	AGVEMVMPGDNVEM	TIELI	YP	PA	IEKGMRFVAREGGRTVGAGVVSIEIIE
FLAFE	AGTEMVMPGDN	NLTVKLI	QPI	IA	MEKGLKFAIREGGRTVGAGQVTEILK
MICLU	EGTEMVMPGD	TTEMSVELI	QPI	IA	MEEGLRFAIREGGRTVGSQVTKITK
MYCGA	EGTEMVMPGDN	TEIIVELIS	SI	ACE	KSGKFSIREGGRTVGAGTVVEVLE
MYCHO	PGREM	VMPGDNVELT	VT	LI	APVIAIEEGTKFSIREGGRTVGAGSVTKILK
MYCLE	EGTEMVMPGDN	TNISVTLI	QPV	AM	DEGLRFAIREGGRTVGAGRVVKIIE
NEIGO	KGVEMVMPGEN	VTTITVELI	APV	IA	MEEGLRFAIREGGRTVGAGVSSVIA
PLAROA	EGTEMVMPGDN	TEMTVQLI	QPI	IA	MEEGLKFAIREGGRTVGAGRVTKILK
PLAROB	EGTEMVMPGDN	TEMTVQLI	QPI	IA	MEEGLKFAIREGGRTVGAGRVTKILK
RICPR	SKDQ	VMPGDNATFS	VELI	KPI	AMQEGKFSIREGGRTVGAGIVTKINN
SALTY	EGVEMVMPGDN	IKMVT	LIHP	IA	MDGLRFAIREGGRTVGAGVVAKVLG
SHEFU	EGVEMVMPGDN	IKMVT	LI	CP	IAMDEGLRFAIREGGRTVGAGVAKIITA
SPIPL	ADDGMV	IPGDR	IN	MTVQLI	CP
STIAU	DNVEMVMPGDN	IAIE	VELIT	P	VAMEKELPFAIREGGRTVGAGVADIIA
STRAU	EGTEMVMPGDN	TDMTVALI	QPV	AM	EEGLKFAIREGGRTVGAGQVTKITK
STRJC	EGTEMVMPGDN	AE	TN	VL	IQPVAMEEGLRFTIREGGRTVGAGQVVKINK
STORR	AGTEMVMPGDN	VTIDVELI	HP	IA	VEQGTTFISIREGGRTVGSQMVTEIEA
TAXOC	EGVEMVMPGDN	VTISVELI	NA	VA	MEKGLRFAIREGGRTVGAGQVTEILD
THEAQ	QGVEMVMPGDN	VTFTVELI	KP	VA	LEEGLRFAIREGGRTVGAGVVTKILE
THEMA	EGVEMVMPGDH	VE	MEI	ELI	YPVAIEKQRFVAREGGRTVGAGVVEVIE
THETH	PGVEMVMPGDN	VTFTVELI	KP	VA	LEEGLRFAIREGGRTVGAGVVTKILE
THICU	KDKEMVMPGDN	V	SITVKLI	APV	IA
TREHY	GDAQ	I	M	P	GDNANLTIELITPIAMEEKQRFaireGGRTVGAGVVKNIIR
UREUR	AGV	D	L	V	M
WOLSU	EGVEMVMPGDN	V	KIN	VELI	APVIAIE
AQAE1	EGVEMVMPGDN	V	LE	VELI	APVIAIE
AQAE2	EGVEMVMPGDN	V	LE	VELI	APVIAIE
BBUR	EGVEMVMPGDN	V	DI	I	VELISSIAMDKNVEFAVAREGGRTVGAGVVKIIE
ECOL12	EGVEMVMPGDN	IKMVT	LIHP	IA	MDGLRFAIREGGRTVGAGVVAKVLG
ECOL11	EGVEMVMPGDN	IKMVT	LIHP	IA	MDGLRFAIREGGRTVGAGVVAKVLG
HINF	EGVEMVMPGDN	IKMT	VSLIHP	IA	MDQGLRFAIREGGRTVGAGVVKIIE
HPYL	EGVEMVMPGDN	V	KITVELIS	P	VALELGTKFAIREGGRTVGAGVVSNIIE

MGEN	ENTEMVLPGDNASITVELIAPIACEKGSKFSIREGGRTVGAGTVTEVLE
MPNEU	ENTEMVLPGDNTSITVELIAPIACEKGSKFSIREGGRTVGAGSVTKCLN
MTUB	EGTEMVMPGDNTNISVKLIQPVAMDEGLRFAIREGGRTVGAGRVTKIK
SYNP3	ADDGMVMPGDRIKMTVELINPIAIEQGMRFaireGGRTIGAGVVSILK
SYNP7	ADDGMVIPGDRIKMTVELINPIAIEQGMRFaireGGRTIGAGVVSILK
TPAL	EGVDMVKPGDNTKIIGELIHPIAMDKGLKLAIREGGRTIASGQVTEILL
GLUPL	IKKEIPSPGENVTCTIHPQKQVVMETLLRFALRDAGRIVGIGAIEARYT
METVA	GEVLEENPGDAAIVKLIPTKPMVIESVLRFAIRDMGMTVAAGMAIQVTA
SULAC	GKEAEKNPGDSAIVKFKPIKELVAEKFLRFAMRDMGKTVGVGVIIDVKP

LIST OF REFERENCES

- Adachi, J. and Hasegawa, M. (1996) MOLPHY version 2.3, programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* 28, 1-150.
- Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H. and Hutchison C. A. (1994) Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl. Acad. Sci. USA* 91, 1569-1573.
- Ahmadian, M. R., Kreutzer, R. and Sprinzl, M. (1991) Overproduction of the *Thermus thermophilus* elongation factor Tu in *Escherichia coli*. *Biochimie* 73, 1037-1043.
- Aimi, J., Badylak, J., Williams, J., Chen, Z. D., Zalkin, H. and Dixon, J. E. (1990) Cloning of a cDNA encoding adenylosuccinate lyase by functional complementation in *Escherichia coli*. *J. Biol. Chem.* 265, 9011-9014.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST, a new generation of protein database search programs. *Nuc. Acids Res.* 25, 3389-3402.
- Andersen, G. R., Pedersen, L., Valente, L., Chatterjee, I., Kinzy, T. G., Kjeldgaard, M. and Nyborg, J. (2000) Structural basis for nucleotide exchange and competition with tRNA in the yeast elongation factor complex eEF1A, eEF1Ba. *Mol. Cell* 6, 1261-1266.
- Arai, K.-I., Kawakita, M. and Kaziro, Y. (1972) Studies on polypeptide elongation factors from *Escherichia coli*. *J. Biol. Chem.* 37(21), 7029-7037.
- Ayala, F. J. (1999) Molecular clock mirages. *Bioessays* 21, 71-75.
- Baldauf, S. L., Palmer, J. D. and Doolittle, W. F. (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* 93(15), 7749-7754.
- Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B. and Steitz, T. A. (1999) Placement of protein and RNA structures into a 5 angstrom-resolution map of the 50S ribosomal subunit. *Nature* 400, 841-847.
- Barbrook, A. C., Lockhart, P. J. and Howe, C. J. (1998) Phylogenetic analysis of plastid origins based on secA sequences. *Curr. Genet.* 34, 336-341.

- Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. L. (2000) The Pfam protein families database *Nucleic Acids Res.* 28, 263-266.
- Benner, S. A. (2001) Natural progression. *Nature* 409, 459.
- Benner, S. A. and Ellington, A. D. (1988) Interpreting the behavior of enzymes. Purpose or pedigree? *CRC Crit. Rev. Biochem.* 23, 369-426.
- Benner, S. A., Ellington, A. D. and Tauer, A. (1989) Modern metabolism as a palimpsest of the RNA world. *Proc. Natl. Acad. Sci. USA* 86(18), 7054-7058.
- Benner, S. A., Cohen, M. A. and Gonnet, G. H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* 229, 1065-1082.
- Benner, S. A., Cannarozzi, G., Gerloff, D., Turcotte, M. and Chelvanayagam, G. (1997) *Bona fide* predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.* 97, 2725-2843.
- Benner, S. A., Trabesinger, N. and Schreiber, D. (1998) Post-genomic science, converting primary structure into physiological function. *Adv. Enzyme Regul.* 38, 155-180.
- Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S. and Knecht, L. (2000) Functional inferences from reconstructed evolutionary biology involving rectified databases--an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* 151, 97 - 106.
- Benner, S. A. and Gaucher, E. A. (2001) Determining function from genomic sequence, A perspective from organic molecular evolutionary geobiochemistry. *Trends Genet.* (submitted).
- Blackburn, P. and Moore, S. (1982) Pancreatic ribonuclease. In *The Enzymes* (3rd Ed.) ed. Boyer, P. D. (Academic Press, New York), vol. 15, pp.317-433.
- Blank, J., Grillenbeck, N. W., Kreutzer, R. and Sprinzl, M. (1995) Overexpression and purification of *Thermus thermophilus* elongation factors G, Tu, and Ts from *Escherichia coli*. *Protein Expression and Purification* 6, 637-645.
- Bork, P. and Koonin, E. V. (1998) Predicting functions from protein sequences - where are the bottlenecks? *Nat. Genet.* 18, 313-318.
- Brown, J. R. and Doolittle, W. F. (1995) Root of the universal tree of life based on ancient aminoacyl-transfer-rna synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* 92(7), 2441-2445.

- Bult, C. J. , White, O. , Olsen, G. J. , Zhou, L. , Fleischmann, R. D. , Sutton, G. G. , Blake, J. A. , FitzGerald, L. M. , Clayton, R. A. , Gocayne, J. D. , Kerlavage, A. R. , Dougherty, B. A. , Tomb, J. F. , Adams, M. D. , Reich, C. I. , Overbeek, R. , Kirkness, E. F. , Weinstock, K. G. , Merrick, J. M. , Glodek, A. , Scott, J. L. , Geoghagen, N. S. M. and Venter, J. C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273(5278), 1058-73.
- Burggraf, S., Olsen, G. J., Stetter, K. O. and Woese, C. R. (1992) A phylogenetic analysis of *Aquifex pyrophilus*. *Syst. Appl. Microbiol.* 15(3), 352-356.
- Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967) Phylogenetic analysis, models and estimation procedures. *Am. J. Hum. Genet.* 19, 233-257.
- Cetin, R., Anborgh, P. H., Cool, R. H. and Parmeggiani, A. (1998) Functional role of the noncatalytic domains of elongation factor Tu in the interactions with ligands. *Biochemistry* 37, 486-495.
- Chandrasekharan, U. M., Sanker, S., Glynias, M. J., Karnik, S. S. and Husain, A. (1996) Angiotensin II forming activity in a reconstructed ancestral chymase. *Science* 271, 502-505.
- Chang B. S. W. and Donoghue, M. J. (2000) Recreating ancestral proteins. *Trends Ecol. Evol.* 15, 109-114.
- Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H. and Benner, S. A. (1997) An analysis of simultaneous variation in protein structures. *Prot. Engineer.* 10, 307-316.
- Chircurel, M. (2000) Whatever happened to leptin? *Nature* 404, 538-540.
- Collins, L. J., Moulton, V. and Penny, D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.* 51, 194-204.
- Cool R. H. and Parmeggiani, A. (1991) Substitution of histidine-84 and the GTPase mechanism of elongation factor-Tu. *Biochemistry* 30(2), 362-366.
- Czworkowski, J. and Moore, P. B. (1996) The elongation phase of protein synthesis. *Prog. Nucleic Acid Res. Mol. Biol.* 54, 293-332.
- Dayhoff, M. O. (1978) Atlas of protein sequences and structure. Vol. 5, Suppl. 3, Natl. Biomed. Res. Found., Washington, DC.
- Delong, E. F., Wu, K. Y., Prezelin, B. B. and Jovine, R. V. M. (1994) High abundance of archaea in antarctic marine picoplankton. *Nature* 371(6499), 695-697.

- Duret, L., Mouchiroud, D. and Gouy, M. (1994) HOVERGEN - A database of homologous vertebrate genes. *Nucleic Acids Res* 22, 2360-2365.
- Duttaroy, A., Bourbeau, D., Wang, X. L. and Wang, E. (1998) Apoptosis rate can be accelerated or decelerated by overexpression or reduction of the level of elongation factor-1 alpha. *Exp. Cell Res.* 238, 168-176.
- Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T.O. (2000) Protein function in the post-genomic era. *Nature* 405, 823-826.
- Fasano, O., Bruns, W., Crechet, J. B., Sander, G. and Parmeggiani, A. (1978) Modification of elongation-factor-Tu/guanine-nucleotide interaction by kirromycin -comparison with effect of aminoacyl transfer RNA and elongation factor-Ts. *Eur. J. Biochem.* 89(2), 557-565.
- Fasano, O. and Parmeggiani, A. (1981) Altered regulation of the guanosine 5'-triphosphatase activity in a kirromycin-resistant elongation factor-Tu. *Biochemistry* 20(5), 1361-1366.
- Fasano, O., De Vendittis, E. and Parmeggiani (1982) Hydrolysis of GTP by elongation factor Tu can be induced by monovalent cations in the absence of other effectors. *J. Biol. Chem.* 257(6), 3145-3150.
- Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27(4), 401-410.
- Felsenstein, J. (1981) Evolutionary trees from DNA-sequences -a maximum-likelihood approach. *J. Mol. Evol.* 17(6), 368-376.
- Felsenstein, J. (1985) Confidence-limits on phylogenies -An approach using the bootstrap. *Evolution* 39(4), 783-791.
- Felsenstein, J. (1988) Phylogenies from molecular sequences, Inference and reliability. *Annu. Rev. Genet.* 22, 521-565.
- Finkelstein, M. K., Fitch, W. M., Lanciani, C. A. and Miyamoto, M. M. (1998) Estimating the probabilities of runs of identical events within biological sequences. *Mol. Biol. Evol.* 15, 470-472.
- Fisher, A., Shi, Y., Ritter, A., Ferretti, J. A., Perez-Lamboy, G., Shah, M., Shiloach, J. and Taniuchi, H. (2000) Functional correlation in amino acid residue mutations of Yeast iso-2-cytochrome *c* that is consistent with the prediction of the concomitantly variable codon theory in cytochrome *c* evolution. *Biochem. Genet.* 38, 177-196.

- Fitch, W. M. (1971) Toward defining the course of evolution, minimum change for a specific tree topology. *Syst. Zool.* 20, 406-416.
- Fitch, W. M. and Margoliash, E. (1967) Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome *c* sequences is of general applicability. *Science* 155, 279-284.
- Fitch, W. M. and Markowitz, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* 4, 579-593.
- Gaidos, E. J., Nealson, K. H. and Kirschvink, J. L. (1999) Biogeochemistry -life in ice-covered oceans. *Science* 284(5420), 1631-1633.
- Galtier, N., Tourasse, N. and Gouy, M. (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science* 283(5399), 220-221.
- Gaucher, E. A., Miyamoto, M. M. and Benner S. A. (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches, Elongation factors. *Proc. Natl. Acad. Sci. USA* 98, 548-552.
- Gerlt, J. A. and Babbitt, P. C. (1998) Mechanistically diverse enzyme superfamilies, The importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* 2, 607-612.
- Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. and Yoshida, M. (1989) Evolution of the vacuolar H⁺-ATPase - implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* 86,661-6665.
- Golding, G. B. (1983) Estimates of DNA and protein-sequence divergence -an examination of some assumptions. *Mol. Biol. Evol.* 1(1), 125-142.
- Golding, G. B. and Dean, A. M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.* 15, 355-369.
- Goldman, N. (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis. *Syst. Zool.* 39, 345-361.
- Gracy, J. and Argos, P. (1998) DOMO, a new database of aligned protein domains. *Trends Biochem. Sci.* 23, 495-497.
- Gribaldo, S. and Cammarano, P. (1998) The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *J. Mol. Evol.* 47(5), 508-516.

- Grishin, N. V., Wolf, Y. I. and Koonin, E. V. (2000) From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.* 10, 991-1000.
- Grosshans, H., Simos, G. and Hurt, E. (2000) Review, transport of tRNA out of the nucleus-direct channeling to the ribosome? *J. Struct. Biol.* 129, 288-294.
- Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664-1674.
- Gu, X. and Li, W.-H. (1998) Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc. Natl. Acad. Sci. USA.* 95, 5899-5905.
- Gupta, R. S. (1998) Protein phylogenies and signature sequences, a reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* 62(4), 1435-1491.
- Gupta, R. S., Bustard, K., Falah, M. and Singh, D. (1997) Sequencing of heat shock protein 70 (dnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. *J. Bacteriol.* 179(2), 345-357.
- Gupta, R. S. and Johari, V. (1998) Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the *Deinococcus-Thermus* group and Cyanobacteria. *J. Mol. Evol.* 46(6), 716-720.
- Harmark, K., Cool, R. H., Clark, B. F. C. and Parmeggiani, A. (1990) The functional and structural roles of residues Gln114 and Glu117 in elongation factor Tu. *Eur. J. Biochem.* 194, 731-737.
- Hasegawa, M., Kishino, H. and Saitou, N. (1991) On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* 32, 443-445.
- Hasegawa, M. and Hashimoto, T. (1993) Ribosomal-RNA trees misleading. *Nature* 361(6407), 23-23.
- Hey, J. (1999) The neutralist, the fly and the selectionist. *Trends Ecol Evol* 14, 35-38.
- Huelsenbeck, J. P., Hillis, D. M. and Jones, R. (1995) Parametric bootstrapping in molecular phylogenetics, Applications and performance. In *Molecular Zoology, Advances, Strategies, and Protocols*, eds. Ferraris, J. and Palumbi, S. (Wiley, New York), pp. 19-45.

- Huelsenbeck, J. P. and Rannala, B. (1997) Phylogenetic methods come of age, testing hypotheses in an evolutionary context. *Science* 276, 227-232.
- Hugenholtz, P., Pitulle, C., Hershberger, K. L. and Pace, N. R. (1998a) Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* 180, (2)366-376.
- Hugenholtz, P., Goebel, B. M. and Pace, N. R. (1998b) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* 180(18), 4765-4774.
- Jain, R., Rivera, M. C. and Lake, J. A. (1999) Horizontal gene transfer among genomes, the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* 96(7), 3801-3806.
- Jermann, T. M., Opitz, J. G., Stackhouse, J. and Benner, S.A. (1995) Reconstructing the evolutionary history of the aridodactyl ribonuclease superfamily. *Nature* 374, 57-59.
- Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS*. 8(3), 275-282.
- Kasting, J. F. (1997) Habitable zones around low mass stars and the search for extraterrestrial life. *Origins of Life and Evolution of the Biosphere* 27(1-3), 291-307.
- Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S. and Leberman, R. (1996) The structure of the *Escherichia coli* EF-Tu/EF-Ts complex at 2.5 angstrom resolution. *Nature* 379, 511-518.
- Kaziro, Y. (1978) Role of guanosine 5'-triphosphate in polypeptide-chain elongation. *Biochim. Biophys. Acta* 505(1), 95-127.
- Klenk, H. P. and Zillig, W. (1994) DNA-dependent RNA-polymerase subunit-b as a tool for phylogenetic reconstructions-branching topology of the archaeal domain. *J. Mol. Evol.* 38(4), 420-432.
- Klenk, H. P., Meier, T. D., Durovic, P., Schwass, V., Lottspeich, F., Dennis, P. P. and Zillig, W. (1999) RNA polymerase of *Aquifex pyrophilus*, implications for the evolution of the bacterial rpoB operon and extremely thermophilic bacteria. *J. Mol. Evol.* 48(5), 528-541.
- Krab, I. M. and Parmeggiani, A. (1998) EF-Tu, a GTPase odyssey. *Biochim. Biophys. Acta* 1443, 1-22.

- Krasny, L., Mesters, J. R., Tieleman, L. N., Kraal, B., Fucik, V., Hilgenfeld, R. and Jonak, J. (1998) Structure and expression of elongation factor Tu from *Bacillus stearothermophilus*. *J. Mol. Biol.* 283(2), 371-81.
- Kumar, S., Tamura, K. and Nei, M. (1993) MEGA, Molecular Evolutionary Genetic Analysis, version 1.0. Pennsylvania State University, University Park, PA.
- Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L. and Pace, N. R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analysis. *Proc. Natl. Acad. Sci. USA* 82, 6955-6959.
- Larget, B. and Simon, D. L. (1999) Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16, 750-759.
- Lechner, K. and Bock, A. (1987) Cloning and nucleotide-sequence of the gene for an archaeobacterial protein-synthesis elongation factor-Tu. *Mol. Gen. Genet.* 208(3), 523-528.
- Lewis, P. O. (2001) Phylogenetics systematics turns over a new leaf. *Trends Ecol. Evol.* 16, 30-37.
- Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150-174.
- Li, W.-H. and Gouy, M. (1991) Statistical methods for testing molecular phylogenies. In *Phylogenetic analysis of DNA sequences* (Miyamoto, M.M. and Cracraft, J, eds), pp. 249-277, Oxford University Press, New York.
- Liberles, D. A., Schreiber, D. R., Govindarajan, S., Chamberlin, S. G. and Benner, S. A. (2001) The Adaptive Evolution Database (TAED). *Genome Biol.* 2, 0003.1-0003.18.
- Lichtarge, O., Bourne, H. R. and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257, 342-358.
- Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* 227, 1435-1441.
- Lockhart, P. J., Steel, M. A., Barbrook, A. C., Huson, D. H., Charleston, M. A. and Howe, C. J. (1998) A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* 15, 1183-1188.
- Lockhart, P. J., Howe, C. J., Barbrook, A. C., Larkum, A. W. D. and Penny, D. (1999) Spectral analysis, systematic bias, and the evolution of chloroplasts. *Mol. Biol. Evol.* 16, 573-576.

- Lockhart, P. J., Huson, D., Maier, U., Fraunholz, M. J., Van de Peer, Y., Barbrook, A. C., Howe, C. J. and Steel, M. A. (2000) How molecules evolve in eubacteria. *Mol. Biol. Evol.* 17, 835-838.
- Lopez, P., Forterre, P. and Philippe, H. (1999) The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* 49, 496-508.
- Ludwig, W., Weizenegger, M., Betzl, D., Leidel, E., Lenz, T., Ludvigsen, A., Mollenhoff, D., Wenzig, P. and Schleifer, K. H. (1990) Complete nucleotide sequences of seven eubacterial genes coding for the elongation factor Tu, functional, structural and phylogenetic evaluations. *Arch. Microbiol.* 153(3), 241-7.
- Maher, K. A. and Stevenson, D. J. (1988) Impact frustration of the origin of life. *Nature* 331(6157), 612-614.
- Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. and Wilson, A. C. (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* 345, 86-89.
- Marin, I., Fares, M. A., Gonzalez-Candela, F., Barrio, E. and Moya, A. (2001) Detecting changes in the functional constraints of paralogous genes. *J. Mol. Evol.* 52, 17-28.
- Masullo, M., Raimo, G., Parente, A., Gambacorta, A., De Rosa, M. and Bocchini, V. (1991) Properties of the elongation factor 1-alpha from the thermoacidophilic archaeobacterium *Sulfolobus solfataricus*. *J. Biochem.* 199, 529-537.
- Masullo, M., De Vendittis, E. and Bocchini, V. (1994) Archaeobacterial elongation factor 1-alpha carries the catalytic site for GTP hydrolysis. *J. Biol. Chem.* 269(32), 20376-20379.
- Masullo, M., Ianniciello, G., Arcari, P. and Bocchini, V. (1997) Properties of truncated forms of the elongation factor 1-alpha from the archaeon *Sulfolobus solfataricus*. *Eur. J. Biochem.* 243, 468-473.
- Messier, W. and Stewart, C. B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* 385, 151-154.
- Miyamoto, M. M. and Fitch, W. M. (1995) Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* 12, 503-513.
- Mojzsis, S. J., Arrhenius, G., McKeegan, K. D., Harrison, T. M., Nutman, A. P. and Friend, C.R.L. (1996) Evidence for life on Earth before 3,800 million years ago. *Nature* 384(6604), 55-59.

- Mooers, A. O. and Holmes, E. C. (2000) The evolution of base composition and phylogenetic inference. *Trends Ecol. Evol.* 15, 365-369.
- Moreira, D., Le Guyader, H. and Philippe, H. (1999) Unusually high evolutionary rate of the elongation factor 1 α genes from the Ciliophora and its impact on the phylogeny of eukaryotes. *Mol. Biol. Evol.* 16, 234-245.
- Morozov, P., Sitnikova, T., Churchill, G., Ayala, F. J. and Rzhetsky, A. (2000) A new method for characterizing replacement rate variation in molecular sequences, Application of the fourier and wavelet models to *Drosophila* and mammalian proteins. *Genetics* 154, 381-395.
- Naylor, G. J. P. and Gerstein, M. (2000) Measuring shifts in function and evolutionary opportunity using variability profiles, a case study of the globins. *J. Mol. Evol.* 51, 223-233.
- Negrutskii, B. S. and El'skaya, A. V. (1998) Eukaryotic translation elongation factor 1- α , structure, expression, functions, and possible role in aminoacyl-tRNA channeling. *Prog. Nucleic Acid Res. Mol. Biol.* 60, 47-78.
- Nissen, P., Kjeldgaard, M., Thirup, S., Polekhina, G., Reshetnikova, L., Clark, B. F. C. and Nyborg, J. (1995) Crystal structure of the ternary complex of Phe-tRNA(Phe), EF-Tu, and a GTP analog. *Science* 270, 1464-1472.
- Nock, S., Grillenbeck, N., Ahmadian, M. R., Ribeiro, S., Kreutzer, R. and Sprinzl, M. (1995) Properties of isolated domains of the elongation factor Tu from *Thermus thermophilus* HB8. *Eur. J. Biochem.* 234, 132-139.
- Olmea, O., Rost, B. and Valencia, A. (1999) Effective use of sequence correlation and conservation in fold recognition *J. Mol. Biol.* 293, 1221-1239.
- Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere. *Science* 276(5313), 734-740.
- Pauling, L. and Zuckerkandl, E. (1963) Chemical paleogenetics, molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.* 17, S9-S16.
- Peltier, M. R., Raley, L. C., Liberles, D. A., Benner, S. A. and Hansen, P. J. (2000) Evolutionary history of the uterine serpins. *J. Exp. Zool. (Mol. Devel. Evol.)* 288, 165-174.
- Penny, D., Hasegawa, M., Waddell, P. J. and Hendy, M. D. (1999) Mammalian evolution, Timing and implications from using the LogDeterminant transform for proteins of differing amino acid composition. *Syst. Biol.* 48, 76-93.

- Peter, M. E., Schirmer, N. K., Reiser, C. O. A. and Sprinzl, M. (1990) Mapping the effector region in *Thermus thermophilus* elongation-factor Tu. *Biochemistry* 29, 2876-2884.
- Philippe, H. and Laurent, J. (1998) How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* 8(6), 616-623.
- Philippe, H. and Forterre, P. (1999) The rooting of the universal tree of life is not reliable. *J. Mol. Biol.* 49, 509-523.
- Philippe, H. and Germot, A. (2000) Phylogeny of eukaryotes based on ribosomal RNA, Long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* 17, 830-834.
- Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M. and Le Guyader, H. (2000) Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. Roy. Soc. Lond. Ser. B* 267, 1213-1221.
- Polekhina, G., Thirup, S., Kjeldgaard, M., Nissen, P., Lippmann, C. and Nyborg, J. (1996) Helix unwinding in the effector region of elongation factor EF-Tu-GDP. *Structure* 4, 1141-1151.
- Raley, L. C. (2000) The evolution of artiodactyl seminal ribonuclease, Connecting in vitro behavior and in vivo function by using experimental paleogenomics. University of Florida (Ph.D. Dissertation).
- Reddy, G. P. V. and Pardee, A. B. (1980) Multi-enzyme complex for metabolic channeling in mammalian DNA-replication. *Proc. Natl. Acad. Sci. USA* 77, 3312-3316.
- Rivera, M. C., Jain, R., Moore, J. E. and Lake, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* 95(11), 6239-6244.
- Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70-percent accuracy. *J. Mol. Biol.* 232, 584-599.
- Rzhetsky, A. and Nei, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* 10(5), 1073-1095.
- Sanangelantoni, A. M., Cammarano, R. and Tiboni, O. (1996) Manipulation of the tuf gene provides clues to the localization of sequence element(s) involved in the thermal stability of *Thermotoga maritima* elongation factor Tu. *Microbiology* 142, 2525-2532.

- SAS Institute Inc. (1988) *SAS/Graph[®] User's Guide* (SAS Institute Inc., Cary, NC), Release 6.03, Ed. 549.
- Schopf, J. W. (1993) Microfossils of the early archaean apex chert -new evidence of the antiquity of life. *Science* 260(5108), 640-646.
- Sekiguchi, Y., Kamagata, Y., Syutsubo, K., Ohashi, A., Harada, H. and Nakamura, K. (1998) Phylogenetic diversity of mesophilic and thermophilic granular sludges determined by 16s rRNA gene analysis. *Microbiol. UK.* 144, 2655-2665.
- Sharp, P. M. and Matassi, G. (1994) Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 4, 851-860.
- Sleep, N. H., Zahnle, K. J., Kasting, J. F. and Morowitz, H. J. (1989) Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* 342(6246), 139-142.
- Song, H. W., Parsons, M. R., Rowsell, S., Leonard, G. and Phillips, S. E. V. (1999) Crystal structure of intact elongation factor EF-Tu from *Escherichia coli* in GDP conformation at 2.05 angstrom resolution. *J. Mol. Biol.* 285, 1245-1256.
- Soucek, J., Hrubá, A., Paluska, E., Chudomel, V., Dostal, J. and Matousek, J. (1983) Immunosuppressive effects of bovine seminal fluid fractions with ribonuclease activity. *Folia Biol. (Praha)*, 29, 250-261.
- Stackhouse, J., Presnell, S. R., McGeehan, Nambiar, K. P. and Benner, S. A. (1990) The ribonuclease from an ancient bovid ruminant. *FEBS Lett.* 262, 104-106.
- Steel, M., Huson, D. and Lockhart, P. J. (2000) Invariable sites models and their use in phylogeny reconstruction. *Syst. Biol.* 49, 225-232.
- Stewart, C. B., Schilling, J. W. and Wilson, A. C. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature (London)* 330, 401-404.
- Sullivan, J., Holsinger, K. E. and Simon, C. (1996) The effect of topology on estimates of among-site rate variation. *J. Mol. Evol.* 42(2), 308-312.
- Sullivan, J., Swofford, D. L. and Naylor, G. J. P. (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* 16, 1347-1356.
- Swofford, D. L. (1998) PAUP*, Phylogenetic Analysis Using Parsimony, and other methods. Version 4.0 beta. Smithsonian Institution, Washington, D.C..

- Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996) Phylogenetic Inference in *Molecular Systematics*, eds. Hillis, D. M., Moritz, C. and Mable, B. K (Sinauer, Sunderland, MA), 2nd Ed., pp. 407-514.
- Tateno, Y., Takezaki, N. and Nei, M. (1994) Relative efficiencies of the maximum likelihood, neighbor joining, and maximum parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* 11(2), 261-277.
- Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* 278, 631-637.
- Thorne, J. L., Kishino, H. and Felsenstein, J. Inching toward reality. An improved likelihood model of sequence evolution. *J. Mol. Evol.* 34, 3-16 (1992).
- Tiboni, O., Sanangelantoni, A. M., Cammarano, P., Cimino, L., Di Pasquale, G. and Sora, S. (1989) Expression in *Escherichia coli* of the *tuf* gene from the extremely thermophilic eubacterium *Thermotoga maritima*, Purification of the Thermotoga elongation factor Tu by thermal denaturation of the mesophilic host cell proteins. *System. Appl. Microbiol.* 12, 127-133.
- Trabesinger-Rüf, N. (1997) Molekularer Darwinismus. Ein neuer Denkansatz zur strukturellen Aufklärung der Immunsuppressivität und Antitumoraktivität der seminalen Ribonuklease des Rindes. E.T.H. University, Switzerland (Ph.D. Dissertation), 123485.
- Tuffley, C. and Steel, M. (1998) Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147, 63-91.
- Voet, D. and Voet, J. G. (1995) *Biochemistry* (John Wiley & Sons, Inc., NY), 2nd Ed.
- Weijland, A. and Parmeggiani, A. (1993) Toward a model for the interaction between elongation factor Tu and the ribosome. *Science* 259, 1311-1314.
- Wittgenstein, L. (1993) *Tractatus Logico-Philosophicus*, German Text with English Translation, N.Y., Routledge.
- Woese, C. R. and Fox, G. E. (1977) Phylogenetic structure of the prokaryotic domain, the primary kingdoms. *Proc. Natl. Acad. Sci. USA* 74, 5088-5090.
- Woese, C. R. (1987) Bacterial evolution. *Microbiol. Rev.* 51(2), 221-271.
- Woese, C. R. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. USA* 95(12), 6854-6859.

- Worix, V. L., Burkhart, W. and Spremulli, L. L. (1995) Cloning, sequence analysis and expression of mammalian mitochondrial protein synthesis elongation factor Tu. *Biochim. Biophys. Acta* 1264, 347-356.
- Yang, F., Demma, M., Warren, V., Dharmawardhane, S. and Condeelis, J. (1990) Identification of an actin-binding protein from *Dictyostelium* as elongation factor 1a. *Nature (London)* 347, 494-496.
- Yang, Z. H. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11(9), 367-372.
- Yang, Z. (1997) PAML, a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 15, 555-556.
- Yang, Z. H., Goldman, N. and Friday, A. (1995a) Maximum likelihood trees from DNA sequences, A peculiar statistical estimation problem. *Syst. Biol.* 44(3), 384-399.
- Yang, Z. H., Kumar, S. and Nei, M. (1995b) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641-1650.
- Yang, Z. H. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences, A Markov chain monte carlo method. *Mol. Biol. Evol.* 14, 717-724.
- Yang, Z. H. and Bielawski, J. P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496-503.
- Yang, Z. H., Nielsen, R., Goldman, N. and Pedersen, A. M. K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431-449.
- Zhang, Y. Y., Proenca, R., Maffei, M., Barone, M., Leopold, L. and Friedman, J. M. (1994) Positional cloning of the mouse obese gene and its human homolog. *Nature* 372, 425-432.

BIOGRAPHICAL SKETCH

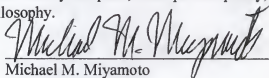
The author, Eric A. Gaucher, was born on January 1, 1972, in Oak Park, Illinois, to Mary and Joseph. He was raised and schooled in the Chicago-land area. From 1990-1992, he attended the liberal arts college Lawrence University in Appleton, Wisconsin. In May of 1994 he graduated from the University of Missouri-Columbia with a Bachelor of Arts degree in biology. In the summer of 1997 he graduated from Loyola University Chicago with a Master of Science degree in Biology under the tutelage of Dr. Howard M. Laten, and then entered the University of Florida's IDP program later that year. Eric recently accepted a post-doc associateship position from the National Research Council/National Aeronautics and Space Administration's Astrobiology Institute.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Steven A. Benner, Chair
Professor of Chemistry

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Michael M. Miyamoto
Professor of Zoology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


Stephen P. Sugrue

Professor of Molecular Cell Biology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Christopher M. West
Associate Professor of Molecular Cell
Biology

This dissertation was submitted to the Graduate Faculty of the College of Medicine and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 2001



Dean, College of Medicine



Dean, Graduate School

UNIVERSITY OF FLORIDA



3 1262 08555 2635